Analyzing data with missing values

Some basic theory and concepts illustrated with simple examples

Problems caused by missing data

- Loss of statistical power
- Bias

Loss of statistical power

Missing values in the data

→ That part of the cannot be used in the analysis

→ **Smaller sample size** in the analysis

→ **Loss of statistical power**

- Not as small effects can be detected (larger p-values)
- Higher uncertainty in the estimates (wider confidence intervals)

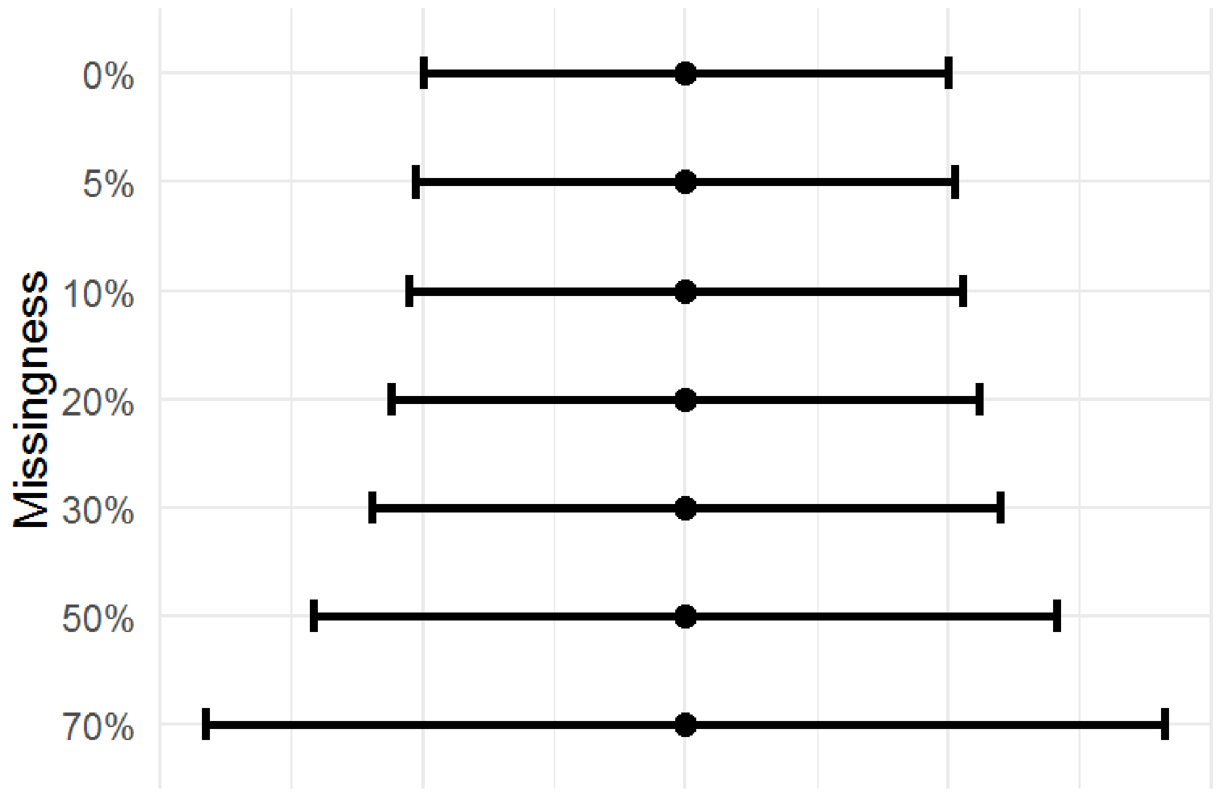
Loss of power (examples)

Original sample size	Missing (%)	Used sample size	(Approx.) factor for general CI width	Smallest correlation (r) *)	Power when r = 0.30 **)	”Average” CI when true r = 0.30
200	0	200	1,00	0,197	0,992	[0.169; 0.422]
200	5	190	1,03	0,202	0,989	[0.165; 0.425]
200	10	180	1,05	0,207	0,985	[0.162; 0.428]
200	20	160	1,12	0,219	0,973	[0.153; 0.436]
200	30	140	1,20	0,234	0,953	[0.142; 0.445]
200	50	100	1,41	0,276	0,865	[0.112; 0.470]
200	75	50	2,00	0,384	0,572	[0.027; 0.536]

- *) Smallest detectable population correlation (r) with α = 0.05 and β = 0.80
- **) Power (β), i.e. probability to observe statistically significant (α = 0.05) correlation when the population correlation r = 0.30

Loss of power (examples)

Missingness	How much wider CIs?
0%	0.0%
5%	2.6%
10%	5.4%
20%	12%
30%	20%
50%	41%
70%	83%



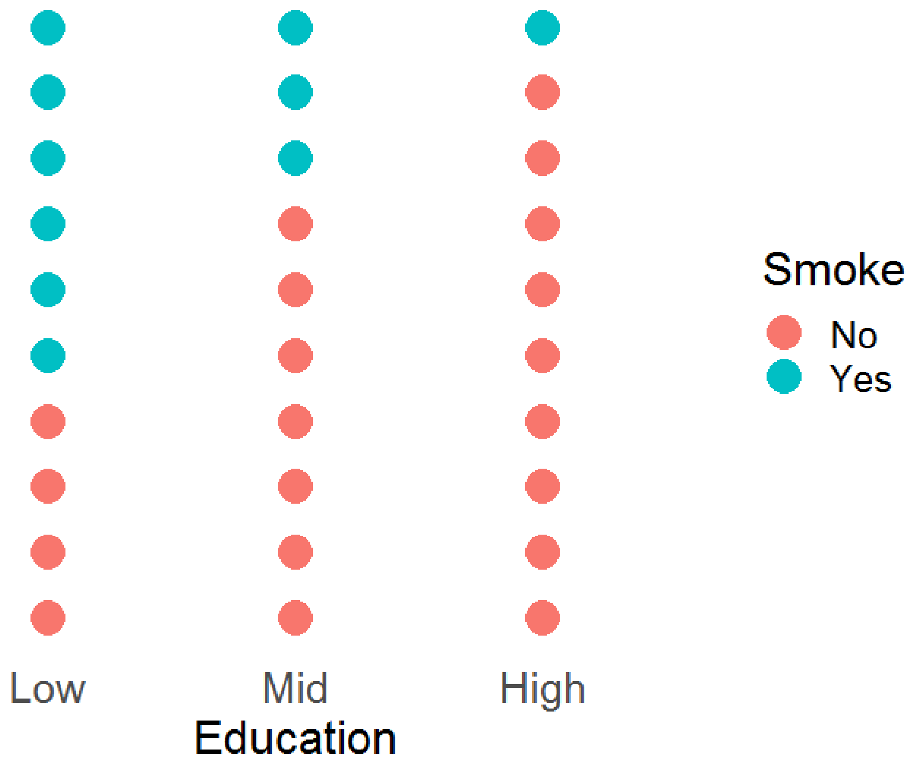
Bias

- **Bias** = The **systematic error** in the estimates
- Missing data causes bias **if** the observed data represents a **subpopulation** where the **association of interest is different** from the association in the whole/target population.
  - Whether there is bias depends also on the **research questions**, not just on the data!

Bias: Example data

Fully observed data

- 30 observations
- Smokers: 10/30 = 33%
  - Low: 6/10 = 60%
  - Mid: 3/10 = 30%
  - High: 1/10 = 10%
- Represents the target population well



Bias: Example 1

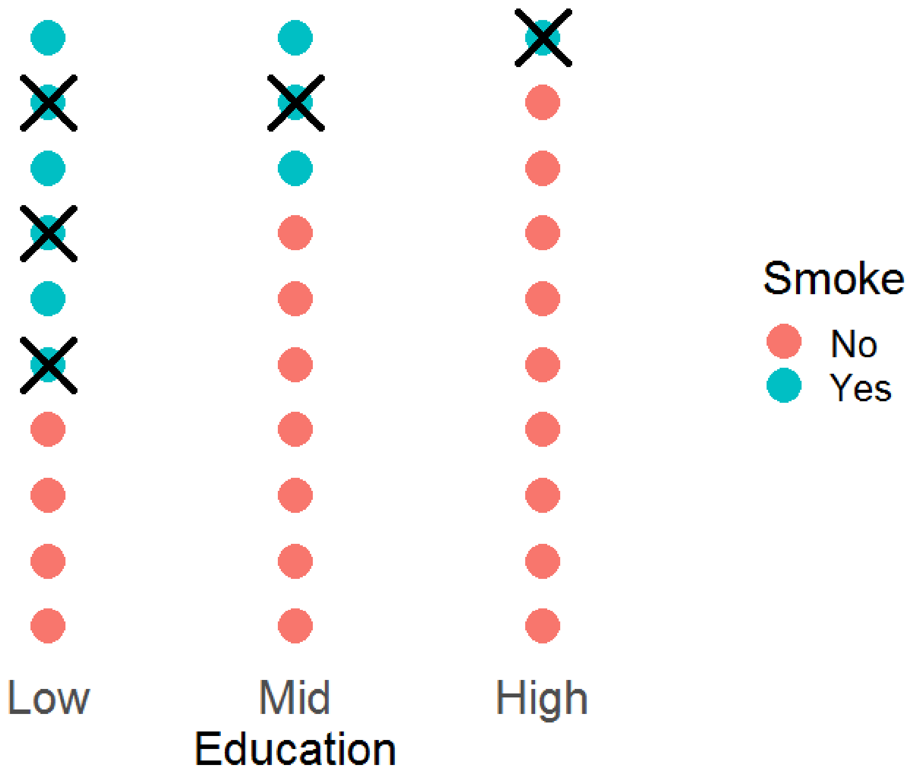
- Research question: What percentage of people smoke?
- Cause of missingness: **Smokers** are more **reluctant to answer** the question about their smoking status.
- Result: **Too low** (i.e. biased) estimate for the percentage of smokers.

Education	Smoking
Low	N/A
High	No
High	No
Mid	N/A
Low	No
High	No
Low	Yes
Mid	No
Mid	No
…	…

Bias: Example 1

Observed data (i.e. the data included in the analysis)

- 25 observations
- Missing data:
    - **50% of the smokers**
- **Smokers: 5/25 = <u>20%</u>**
- The observed data represent a population with smaller proportion of smokers



Bias: Example 2

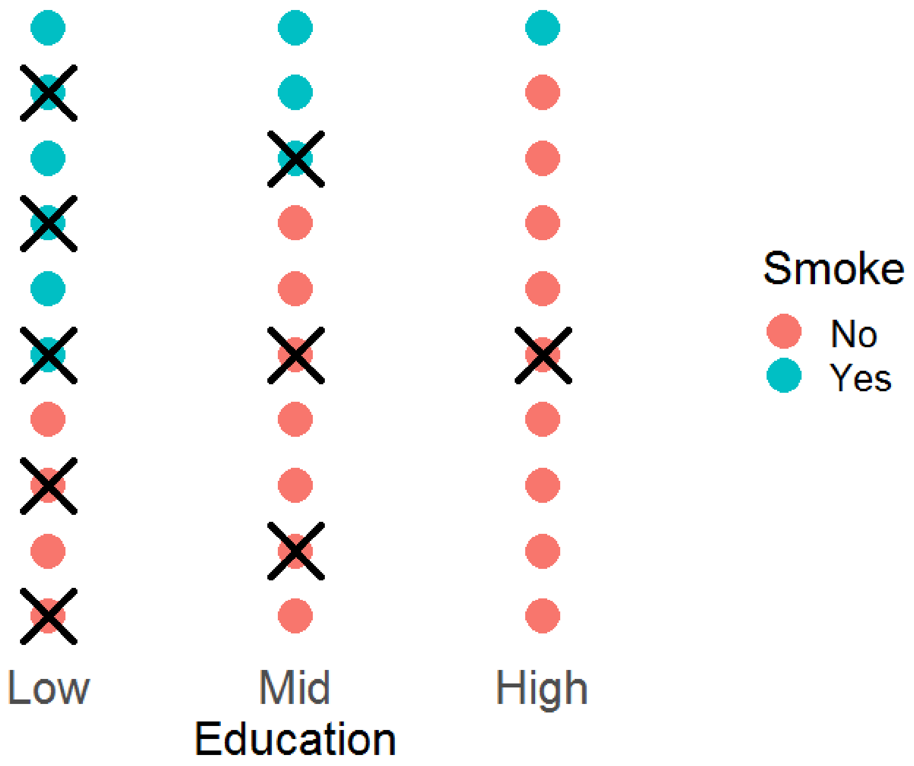
- Research question: What percentage of people smoke?
- Cause of missingness: **Less educated people** are more reluctant to answer.
- Result: **Too low** (i.e. biased) estimate for the percentage of smokers.
  - Reason: The observed data represents a population with higher average education level (than the true education level) and higher educated people smoke less.
  - Education is not included in the analysis model.

Education	Smoking
Low	N/A
High	No
High	No
Mid	Yes
Low	N/A
High	No
Low	Yes
Mid	N/A
Mid	No
…	…

Bias: Example 2

Observed data

- 21 observations
- Missing data:
  - 50% of the low-educated
  - 30% of the mid-educated
  - 10% of the highly educated
- **Smokers: 6/21 = <u>29%</u>**
- Observed data represent a higher educated population (and they smoke less)



Bias: Example 3

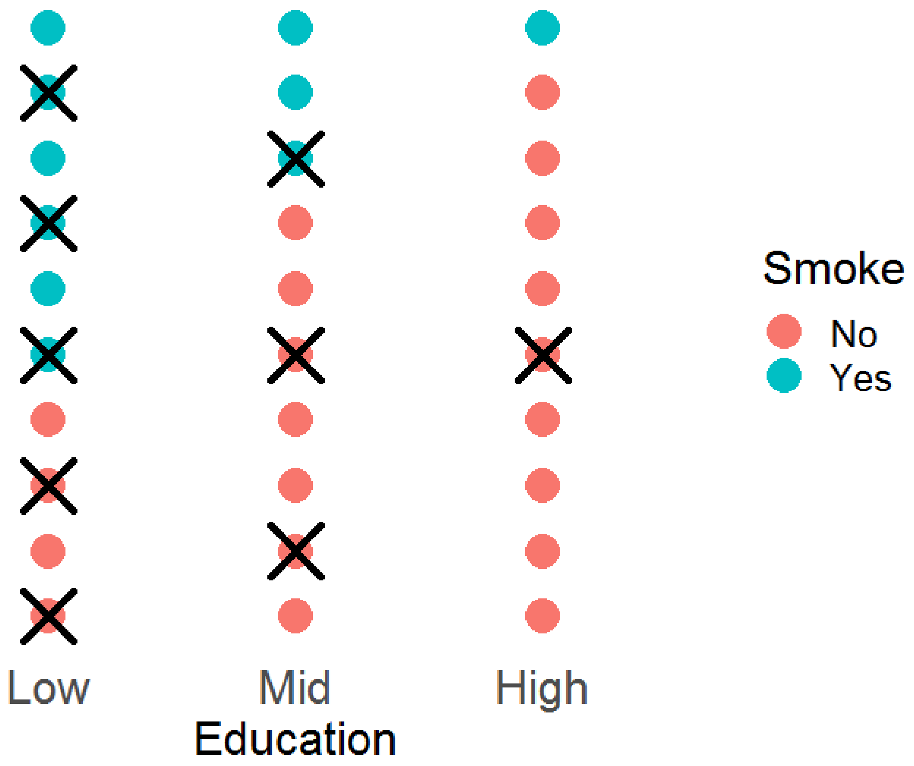
- Research question: How is smoking status **associated with the level of education**?
- Cause of missingness: **Less educated people** are more reluctant to answer.
  - The same as in Example 2!
- Result: **Unbiased** estimates!
  - There is just less data on less educated people but proportions of smokers are unbiased **within each level of education.**
  - From another point of view: There is too small percentage of smokers in the data but because missingness depends only on the education level, and **education level is** (controlled for) **in the analysis model**, the missing data does not cause bias!

Education	Smoking
Low	N/A
High	No
High	No
Mid	Yes
Low	N/A
High	No
Low	Yes
Mid	N/A
Mid	No
…	…

Bias: Example 3

Observed data (the same as in example 2!)

- 21 observations
- Smokers by education level:
  - **Low: 3/5 = 60%**
  - **Mid: 2/7 = 29% ~ 30%**
  - **High: 1/9 = 11% ~ 10%**
- **Unbiased** estimates



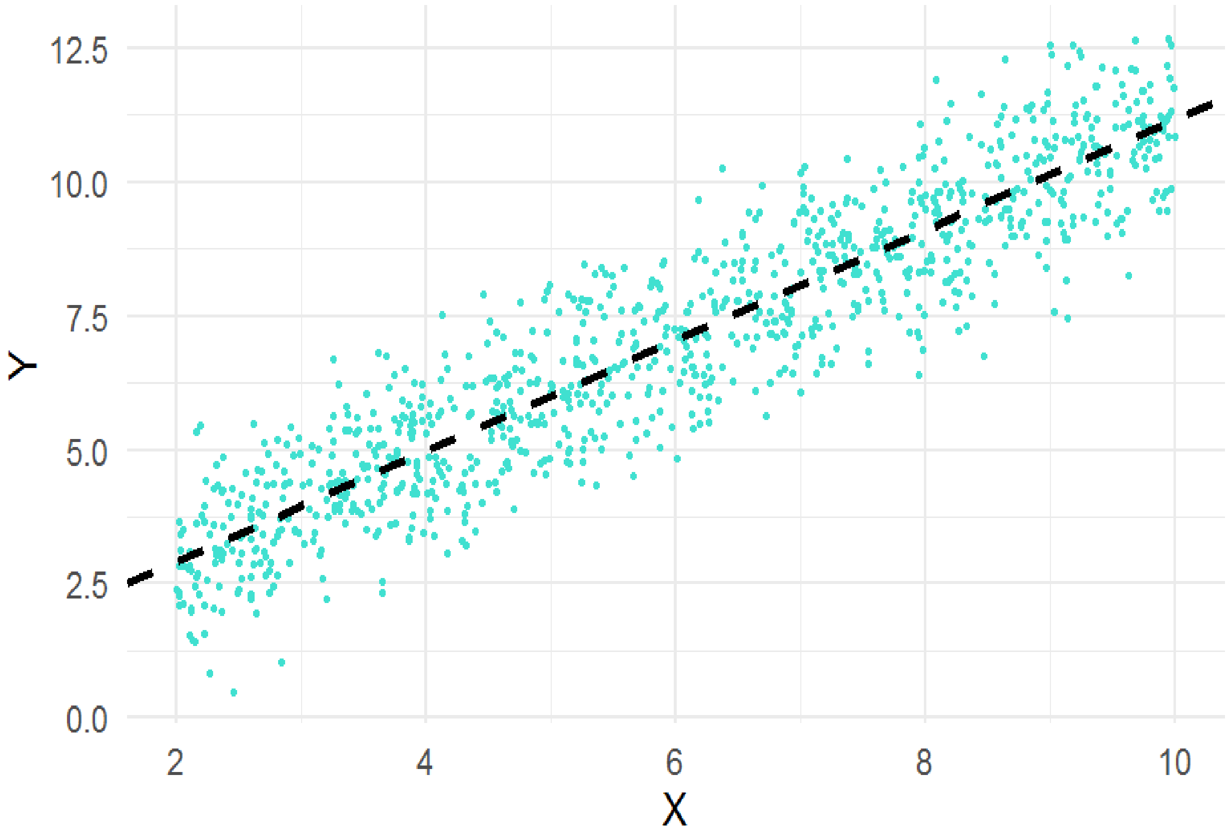
Bias: Example 4

- Research question: How is the **speed of the car (the predictor)** at an accident related to the **severity (1-10) of the driver's injury (the response)?**
- Data: The speed of the car is found out by asking the driver about it. Information about the injuries from everyone.
- Cause of missingness: **Most severely injured drivers cannot answer** the question about their speed.
  - The missignsess **is** in the **predictor** but it **depends** on the **response**!
- Result: **Biased** estimates
  - See Example 5.

Injury	Speed
4	55
6	80
2	40
5	70
8	N/A
2	50
7	60
9	N/A
10	N/A
…	…

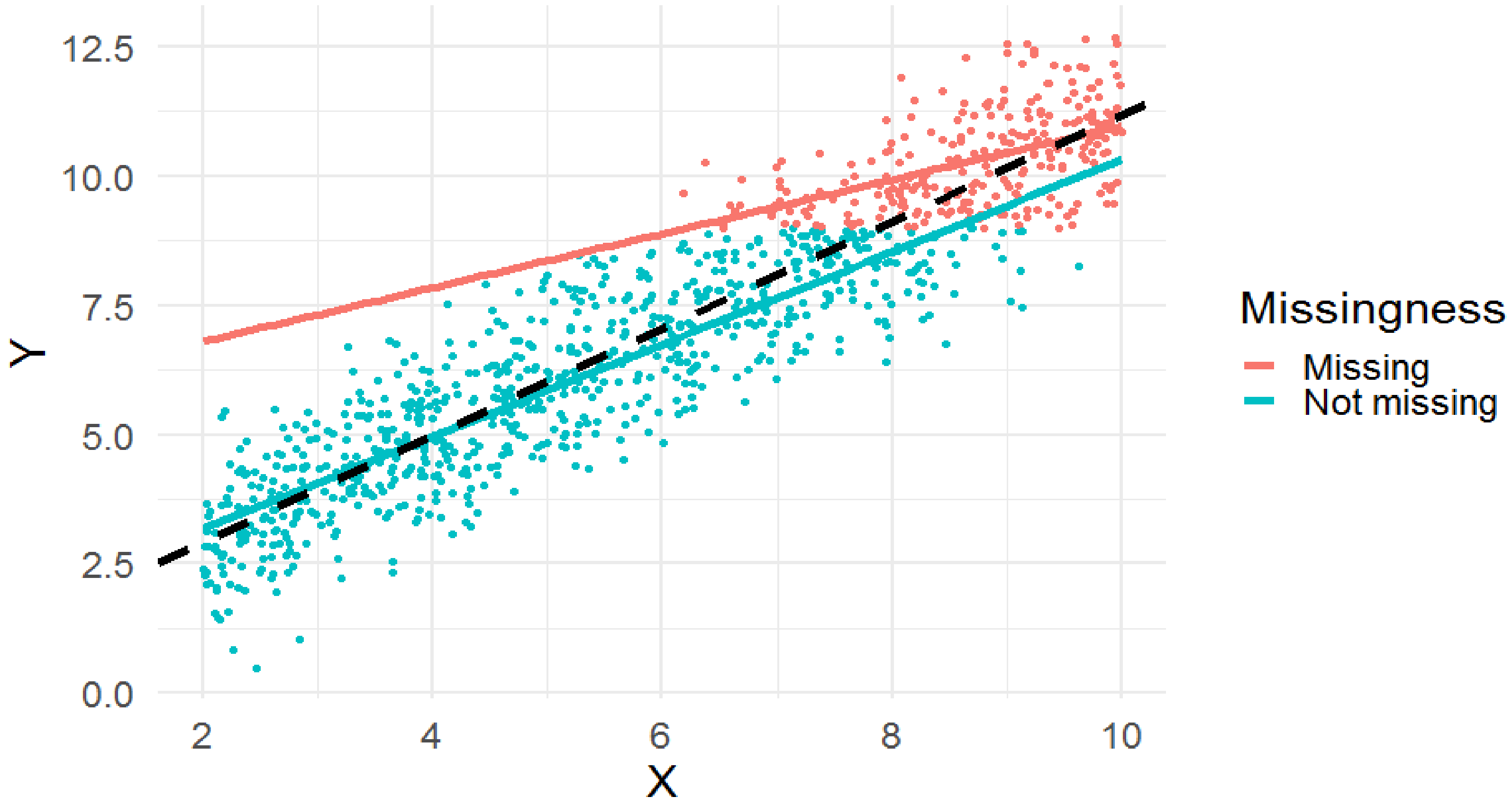
Bias: Examples with continuous data

No missing data



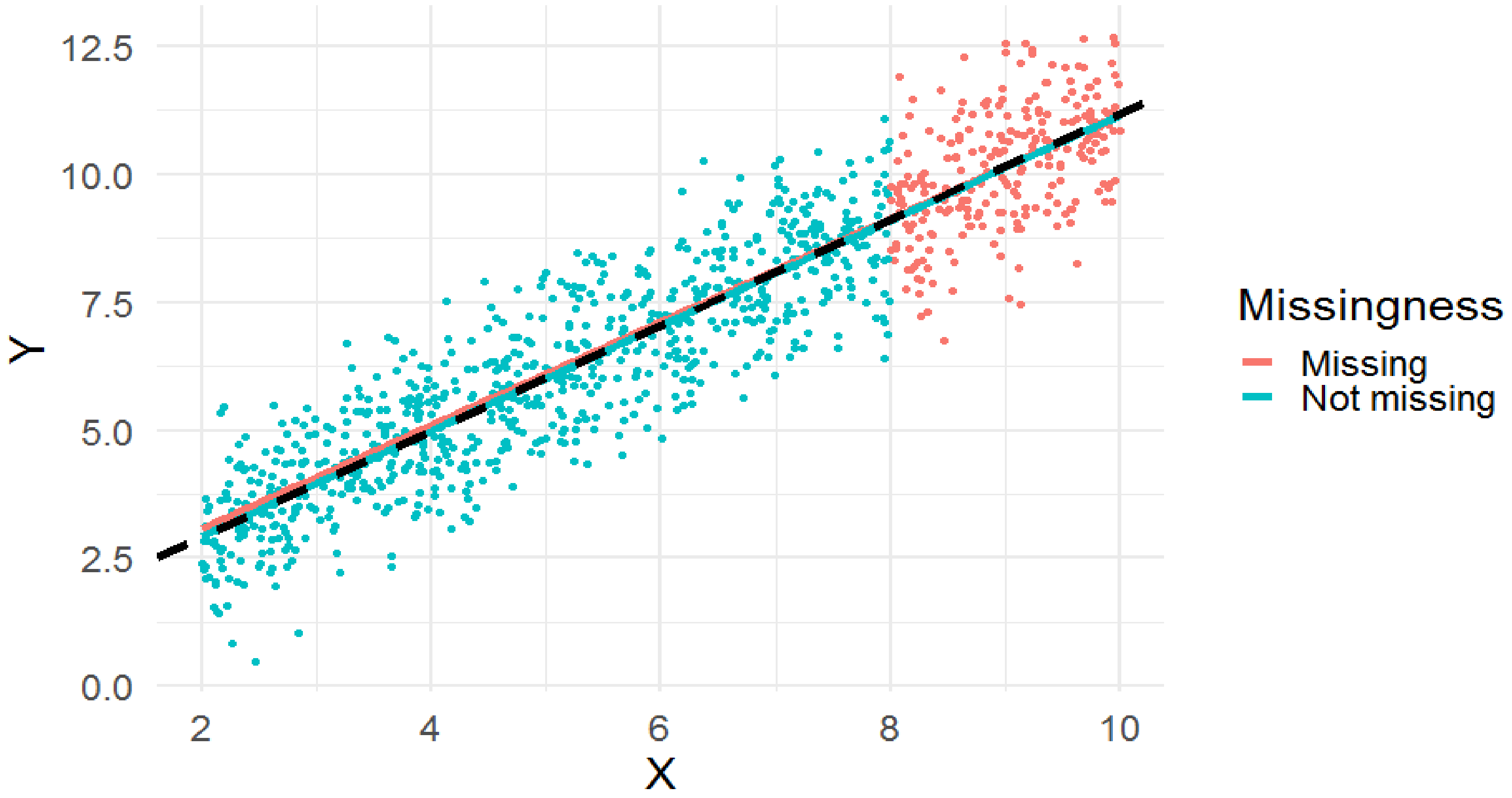
Bias: Example 5

Missingness **<u>depends</u> only on Y. A problem.**



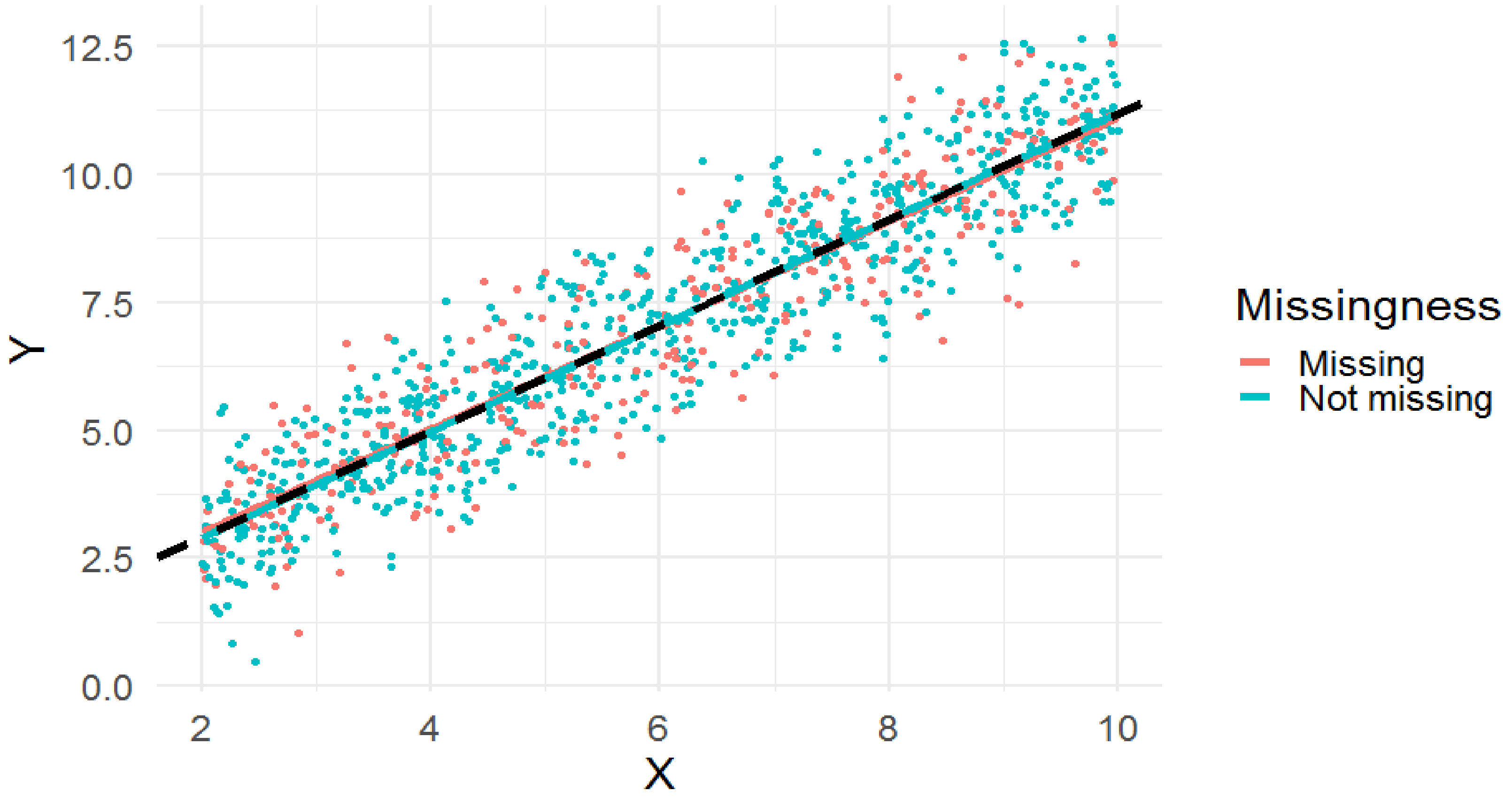
Bias: Example 6

Missingness <u>**depends**</u> **only on X. NOT a problem!**



Bias: Example 7

Missingness **depends <u>neither</u> on X <u>nor</u> on Y. NOT a problem.**



Marginal and conditional (in)dependency

- **Marginal** dependency: Dependency on a variable
- **Conditional** dependency: Dependency on a **variable after the dependency on the other variables "is taken into account"**
- In Examples 3 and 6
  - (Probability of) missingness **does depend** on smoking/Y (when the dependency on Education/X is **<u>not</u>** taken into account) = **Marginal dependency** on Y
  - Missingness does **<u>not</u> depend** on Y when the dependency on X **<u>is</u>** taken into account = **Conditional <u>in</u>dependency** on Y (given X)
- In Example 4 missingness depends marginally, but not conditionally (given injury), on the speed (the predictor).

Bias: Conclusion

- Whether there appears bias or not, depends also on the **research question/analysis model**, not just on the data!
  - E.g. Example 2 vs. Example 3: Bias vs. no bias even though the (observed) data is the same.
- If missingness **depends** only on the **predictors** (i.e. conditional independence on Y) then **no bias** appears!
  - Examples 3 and 6 (and 7)
- Bias appears in Examples 1, 2, 4 and 5 where the missingness is **not conditionally independent of Y** given X

Bias: Notes

- In practice we do not (usually) know the cause/mechanism of missingness but it has to be **assumed**
  - E.g in Example 1 we cannot know, based on the observed data, whether the missingness depends on smoking status or education

Imputation

- Assumptions and purpose
- Methods and their performance

Imputation

- "Imputation is the process of replacing missing data with substituted values" (Wikipedia)
- The loss power and bias caused by missing data can possibly be decreased using imputation if
  - Certain assumptions hold
  - Imputation is done appropriately
- The primary purpose of imputation should <u>not</u> be so much to replace the missing values by as "correct" values as possible, but to get as "correct" **results** as possible from the **analysis**

Imputation: Assumptions

- The missingness in a variable is **conditionally independent of the missing data**, given the observed data
  - i.e. often in practice: The missingness **does not depend on the (imputed) variable(s) itself** when the dependency (of the missingness) on the other variables is taken into account
  - Even more roughly: The data includes the information about missingness
- The **imputation model** is specified “correctly”
  - Imputation model = The (statistical) model that is used to predict the imputed values
  - The relations between the variables need to be modeled/retained

Single imputation

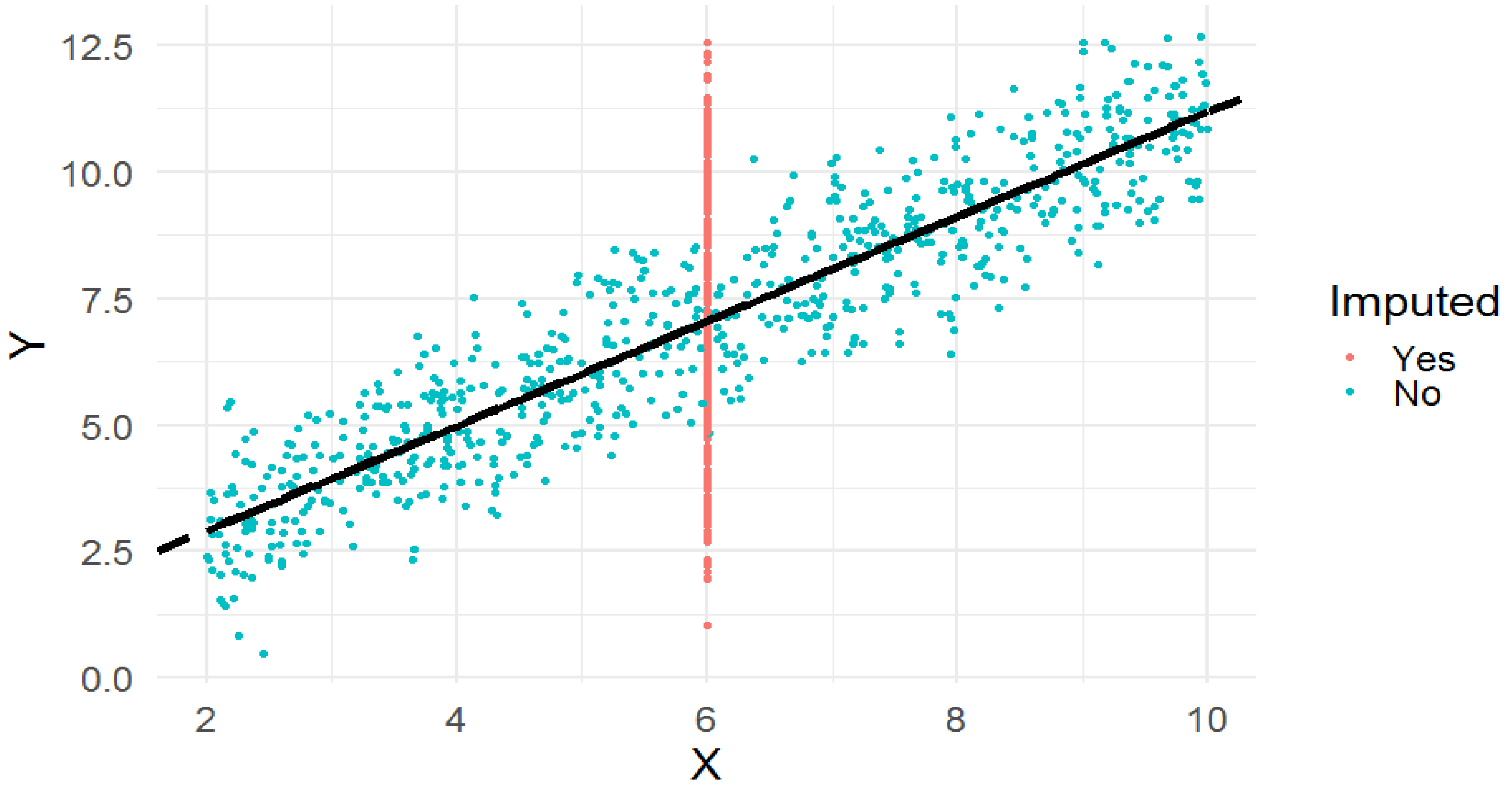
- Missing values are imputed into the data.
- The imputed data are then used in the analyses in a normal way.

Mean/median/mode imputation

- Missing values are replaced by the mean/median/mode of the variable
- Does **not** take into account the **relations between** the variables!
- May **distort badly** the **distributions** of the imputed variables and their **relations** to the other variables!
- Maybe ok if
  - Only small percentage of values are missing
  - The imputed variable(s) are not strongly related to the other variables
  - The imputed variables are not the variables of the main interest

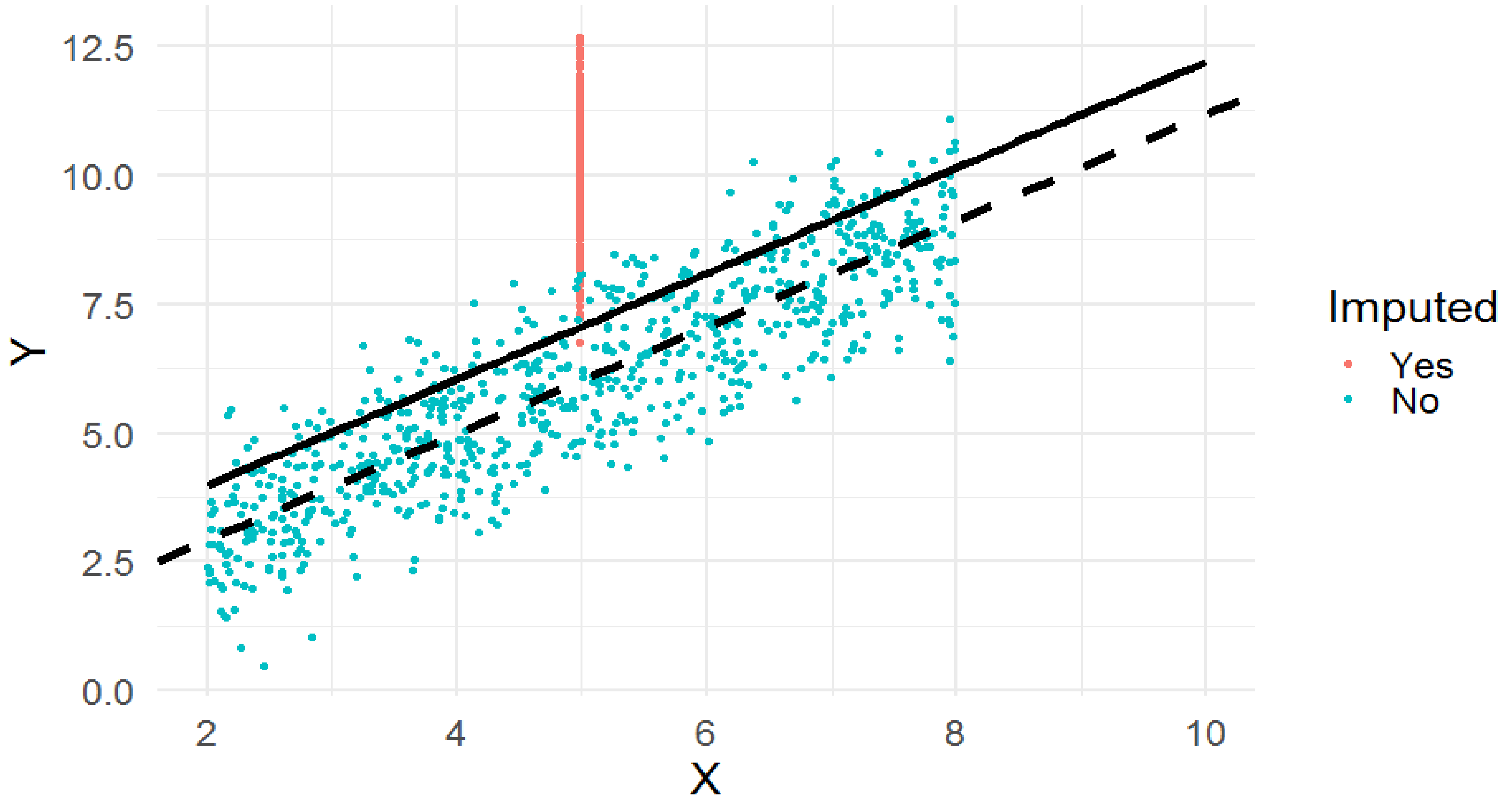
Examples: Mean imputation

Example 7, X missing: No bias but too wide CI



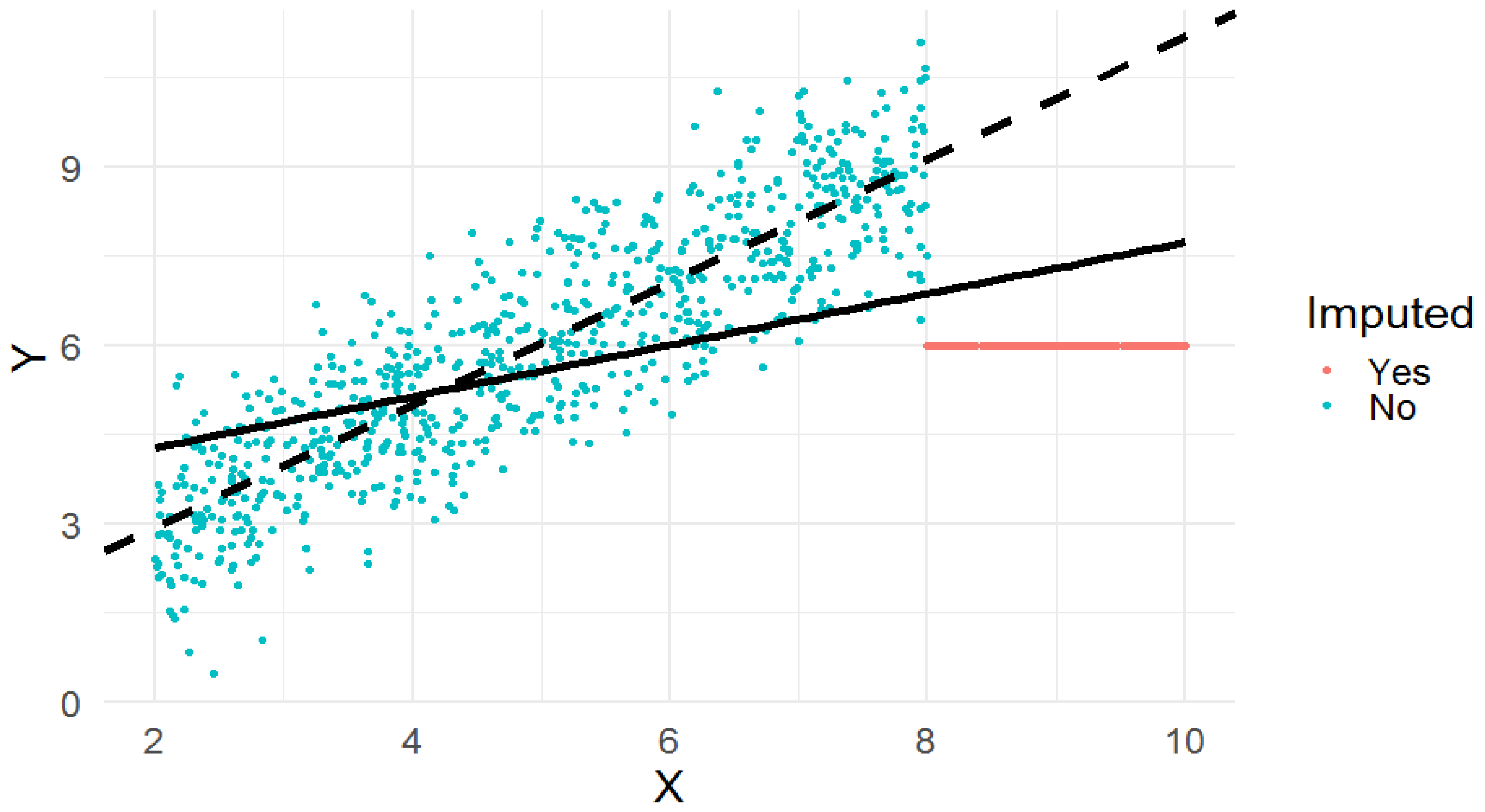
Examples: Mean imputation

Example 6, X missing: Bias in intercept, not in slope, too wide CI



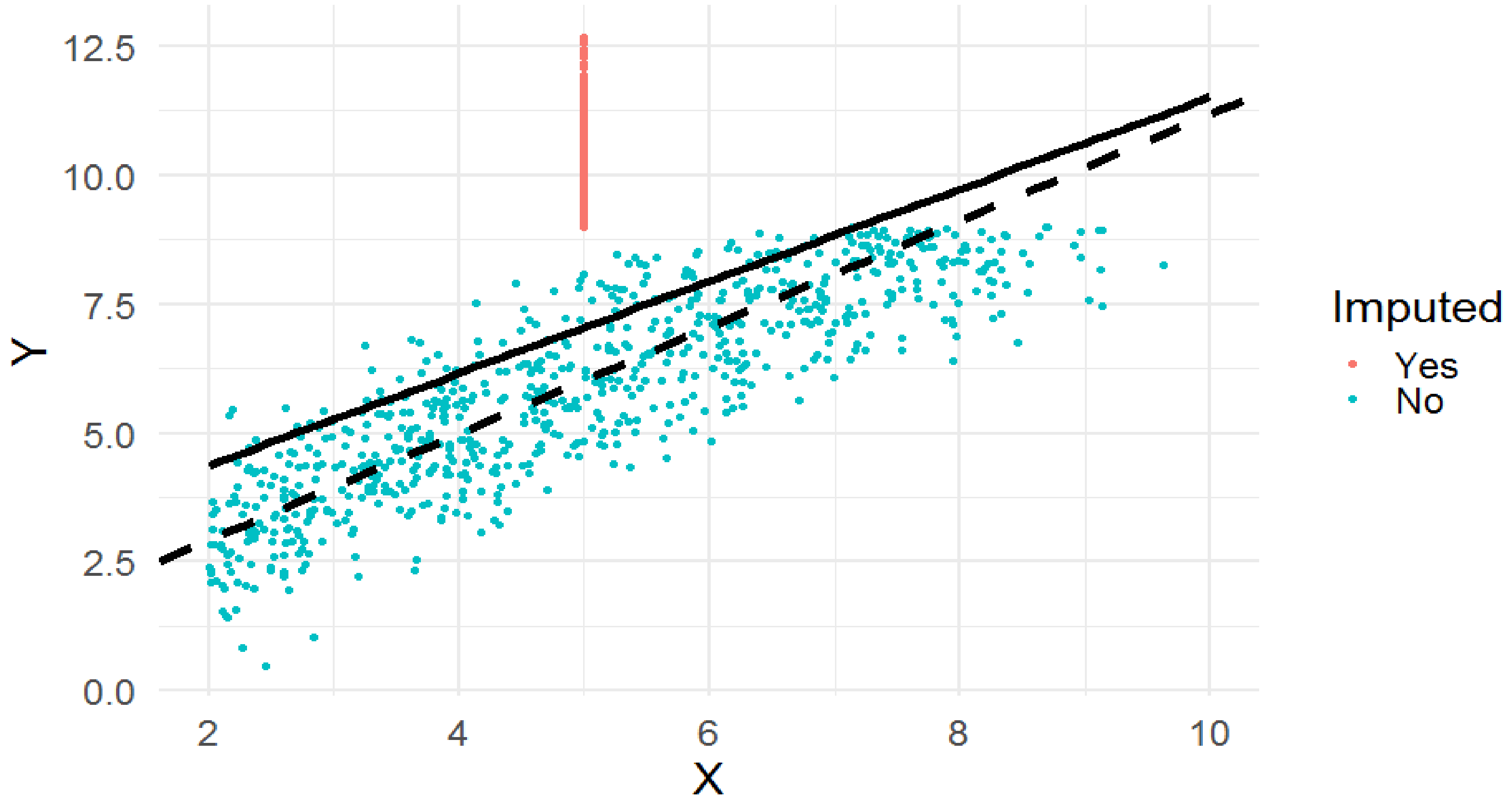
Examples: Mean imputation

Example 6, Y missing: Severe bias!



Examples: Mean imputation

Example 5, X missing: Biased estimates

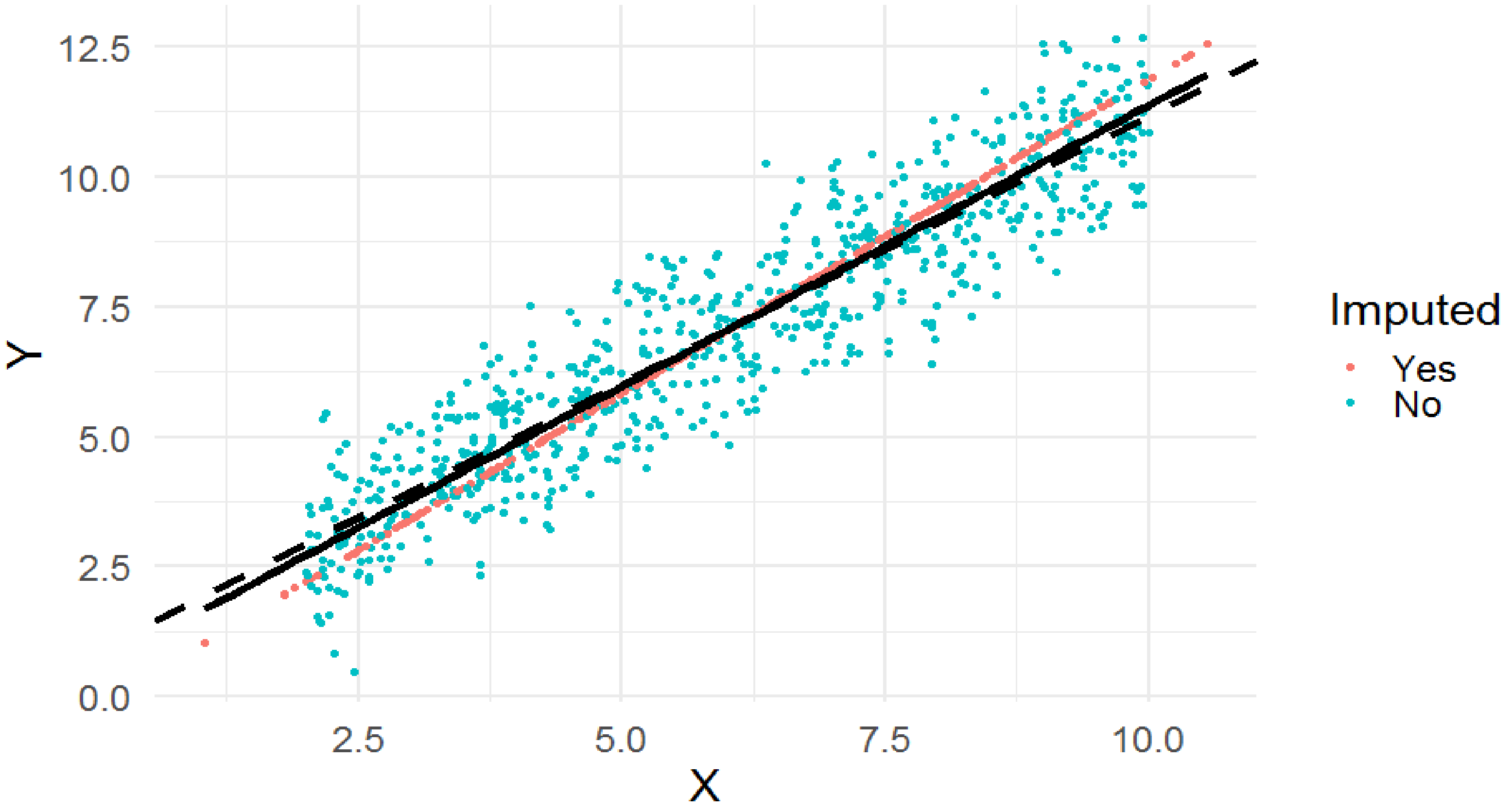


“Regression” methods

- The **other variables are used**, too
  - The substituted values are predicted by e.g. linear regression, logistic regression, regression tree, random forest
- The relations between the variables are retained
  - All relevant variables (including interactions and non-linearities) should be included in the imputation model
- Problem: The associations between the variables are **strengthened** artificially, i.e. **too little variation** in the data
  - Causes too narrow confidence intervals and too small p-values

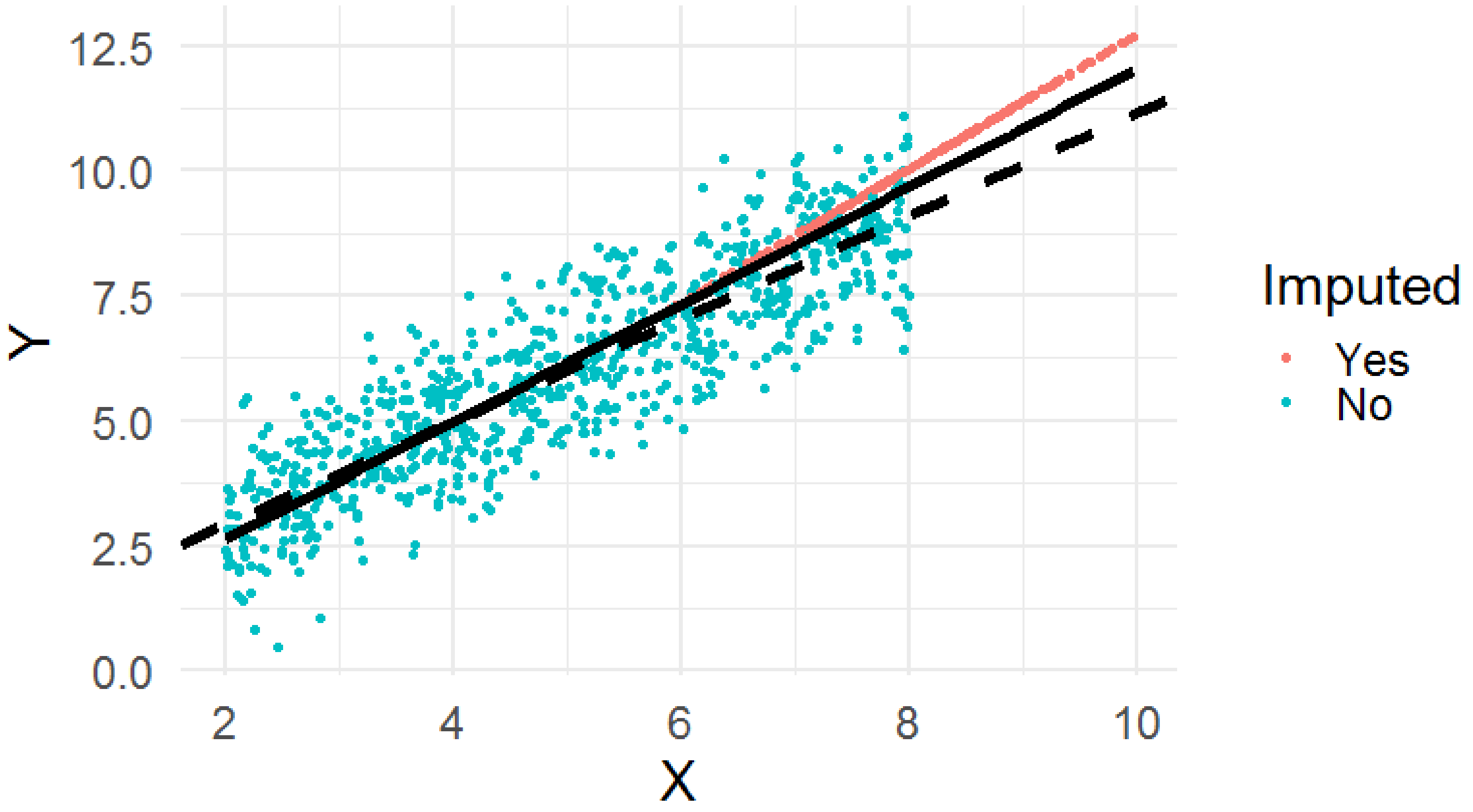
Examples: Regression imputation

Example 7, X missing: Small bias(?), a slightly too narrow CI



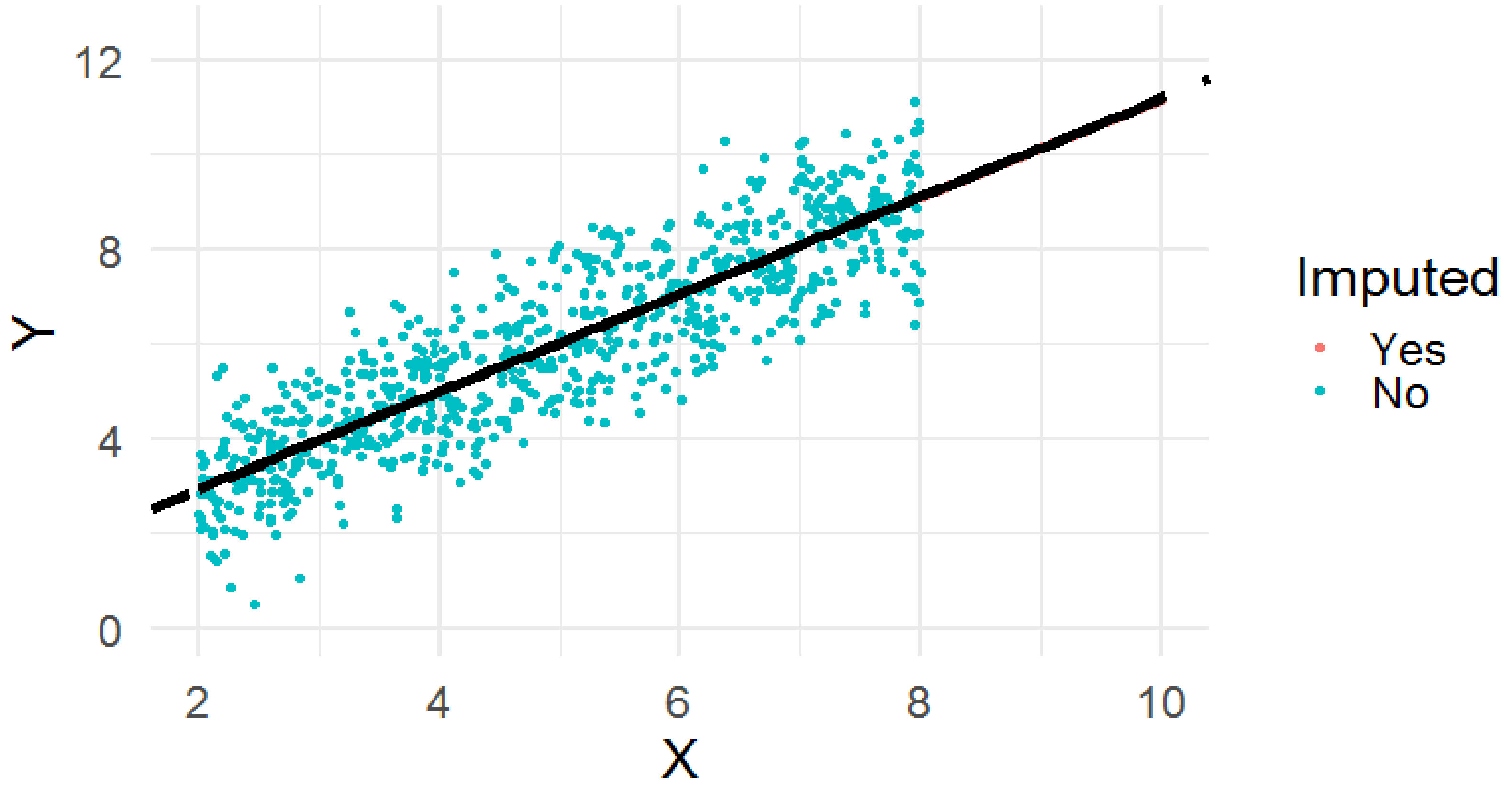
Examples: Regression imputation

Example 6, X missing: Biased estimates



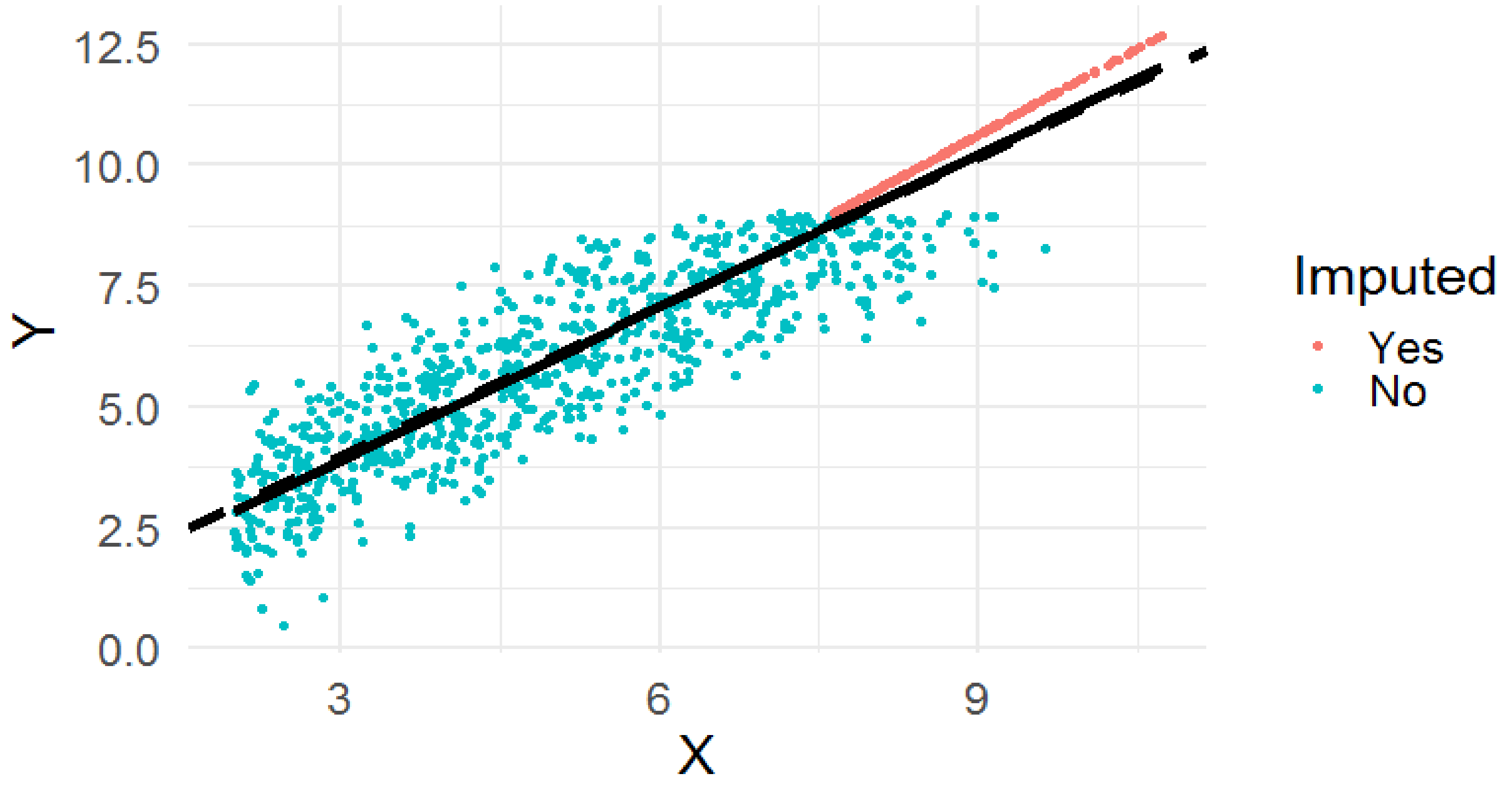
Examples: Regression imputation

Example 6, Y missing: No bias but too narrow CIs



Examples: Regression imputation

Example 5, X missing: Small bias(?), slightly too narrow CI

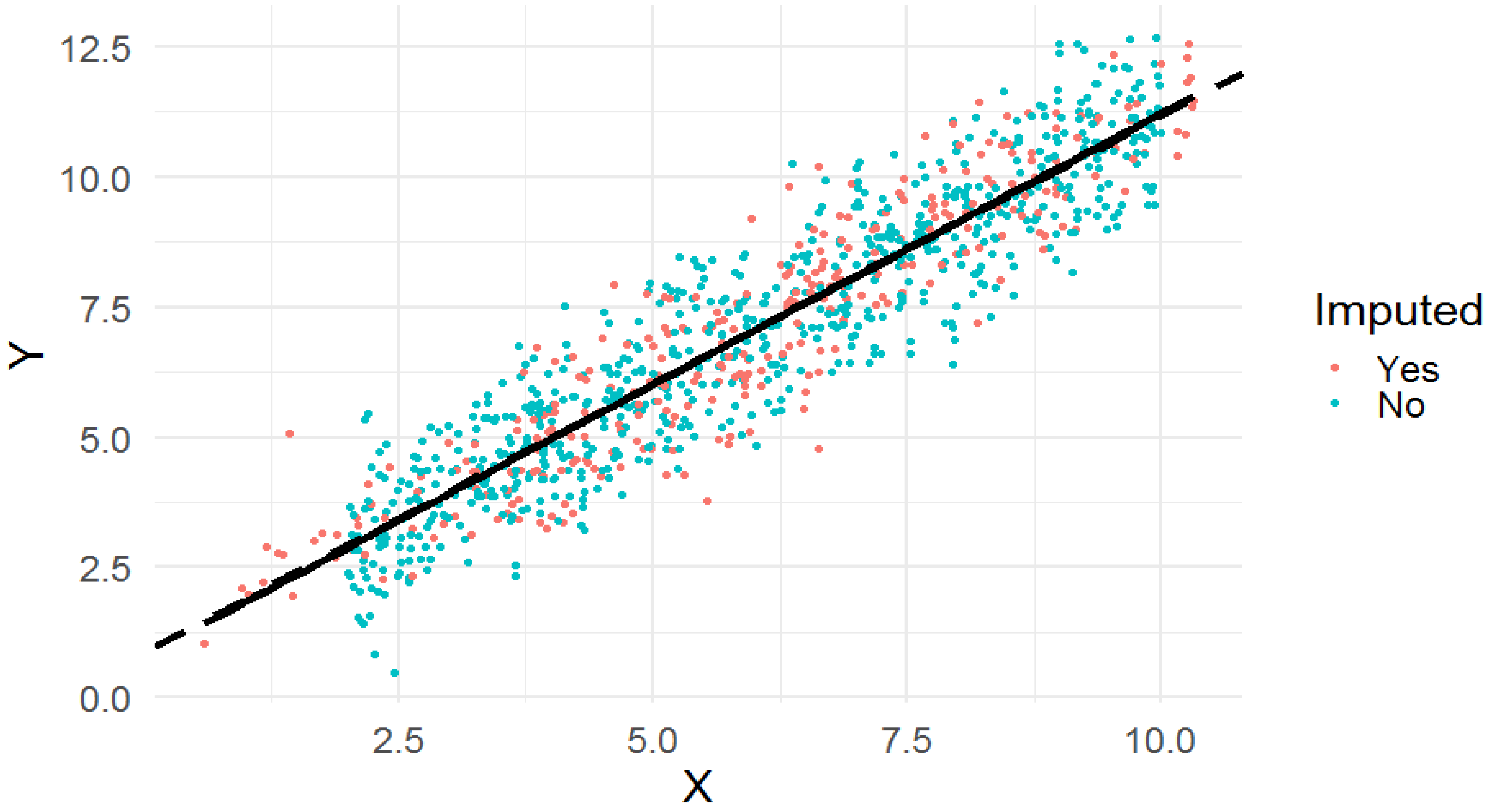


Regression + added variation

- The imputed values consist of
  - values predicted by some regression method
  - added **random error**
- The relations between the variables are retained and the **variation** in the data is “correct”
- There is **uncertainty** in the parameter estimates of the **imputation model** that is not taken into account
  - Still a little too narrow confidence intervals and too small p-values

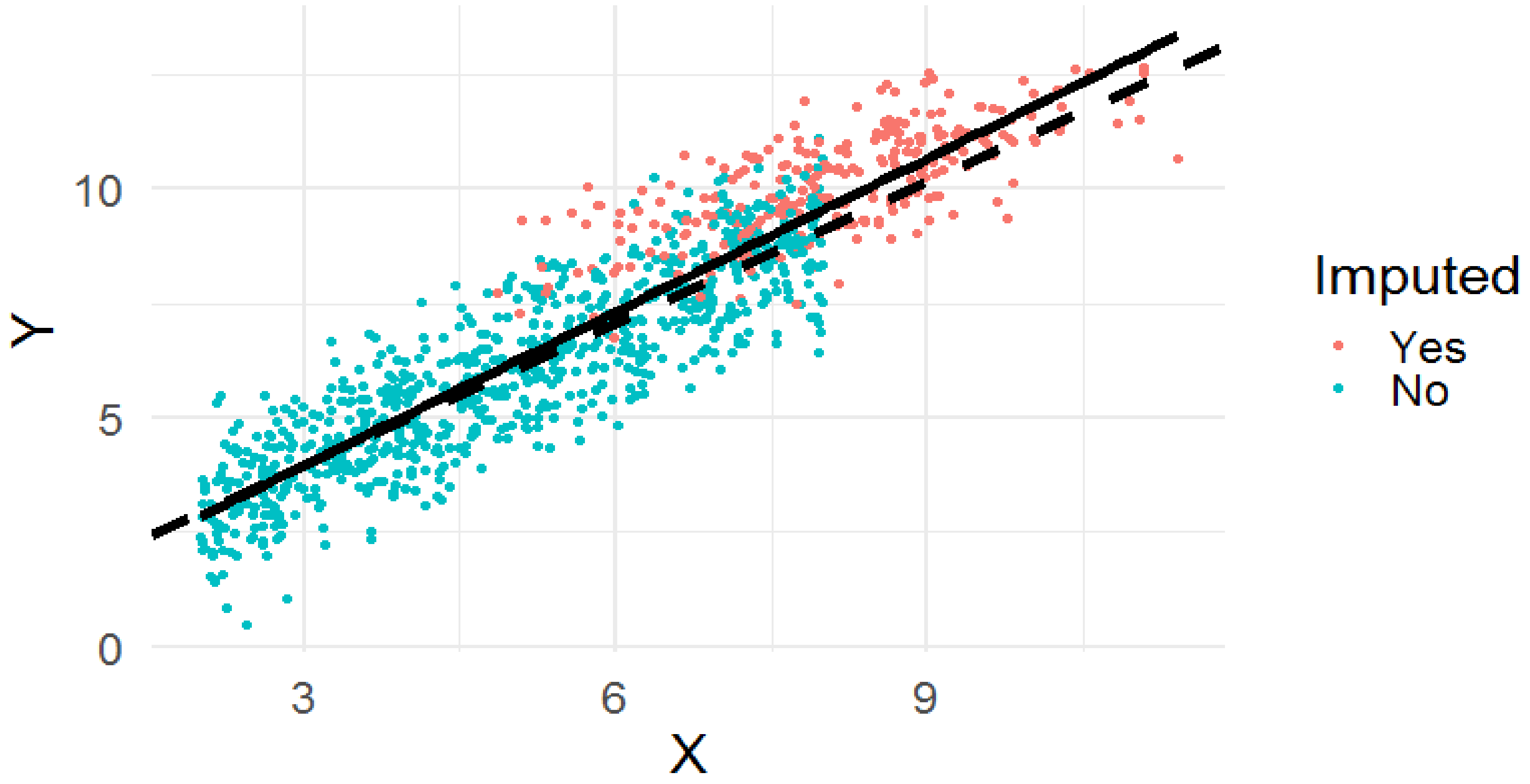
Examples: Regression + added variation

Example 7, X missing: No bias(!), possibly slightly too narrow CI



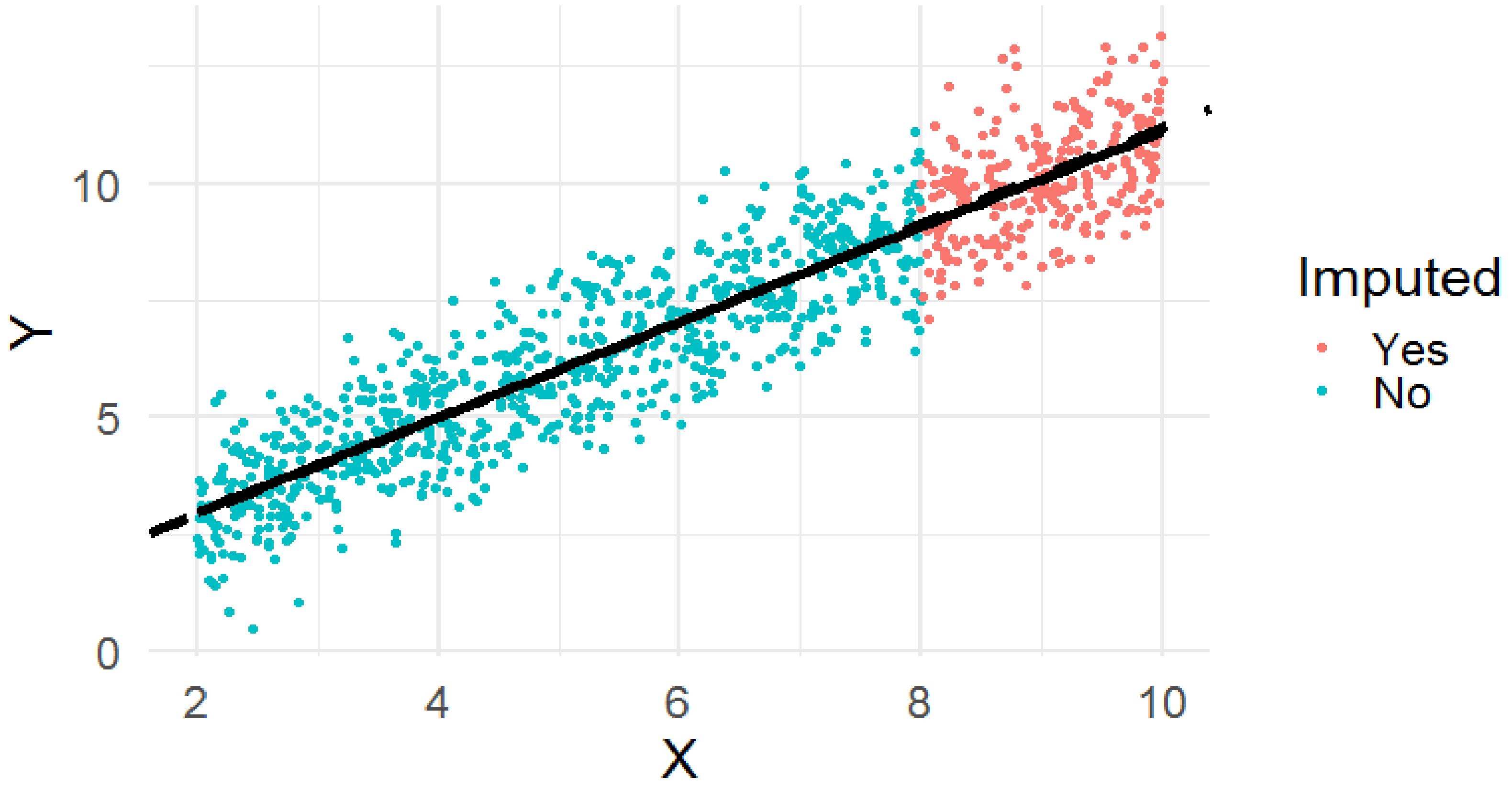
Examples: Regression + added variation

Example 6, X missing: Biased estimates



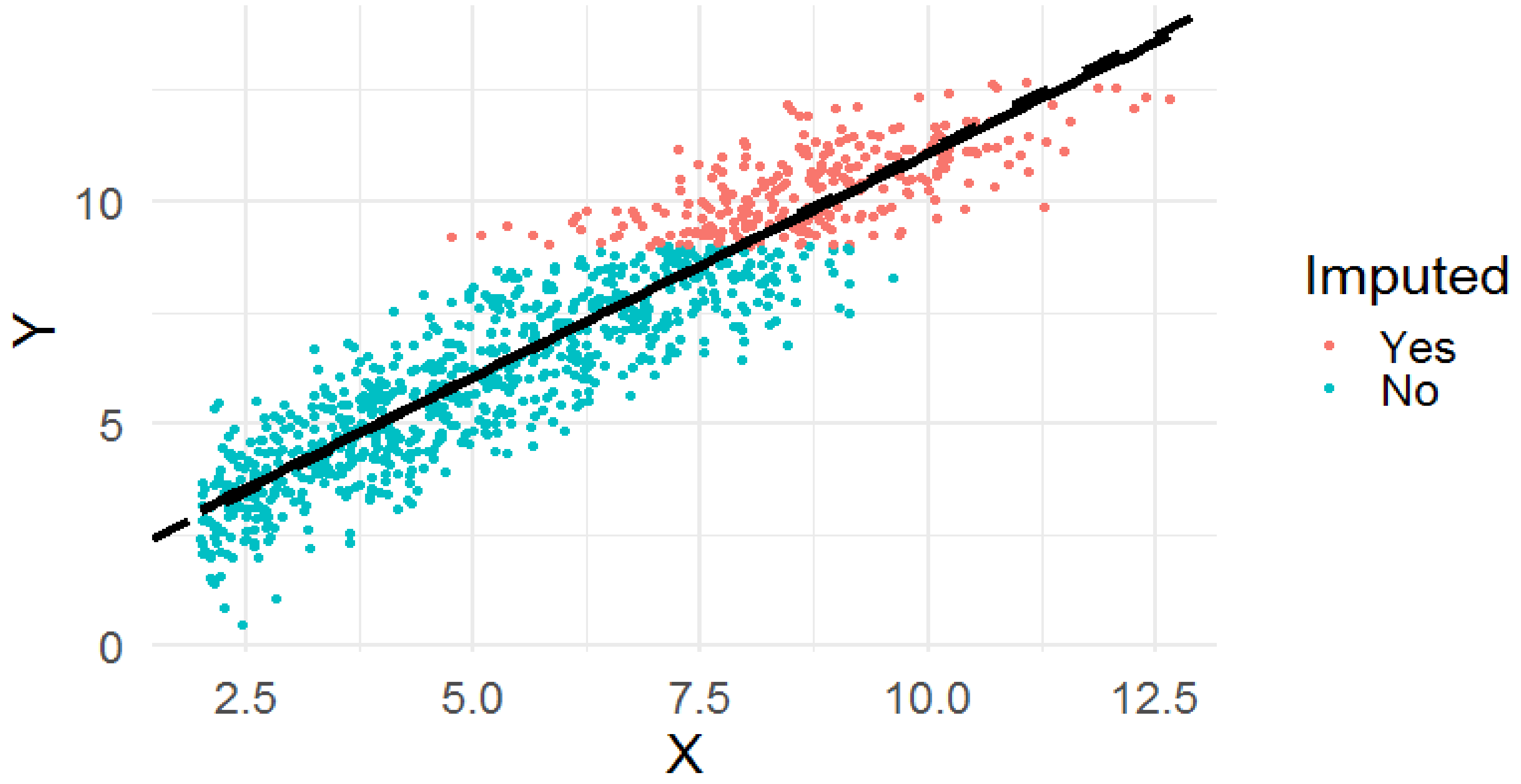
Examples: Regression + added variation

Example 6, Y missing: No bias, “correct” CIs(?)



Examples: Regression + added variation

Example 5, X missing: Tiny bias(?)



Multiple imputation

- **Multiple** imputed datasets are created
  - Some “regression” methods are usually used to predict the imputed values
  - Randomness is “added” to the
    - **Parameters of the imputation model**
    - The values predicted by the imputation model
- The **analysis model** is fitted to **all imputed datasets**
- The results of the multiple analyses are **pooled** (“combined”) to get the final results (Rubin’s rules)
  - The uncertainty in the parameters of the imputation model values is explicit and it is **taken into account when CIs and p-values are calculated**!

Multiple imputation

If the assumptions hold and the imputation model is specified appropriately multiple imputation should give

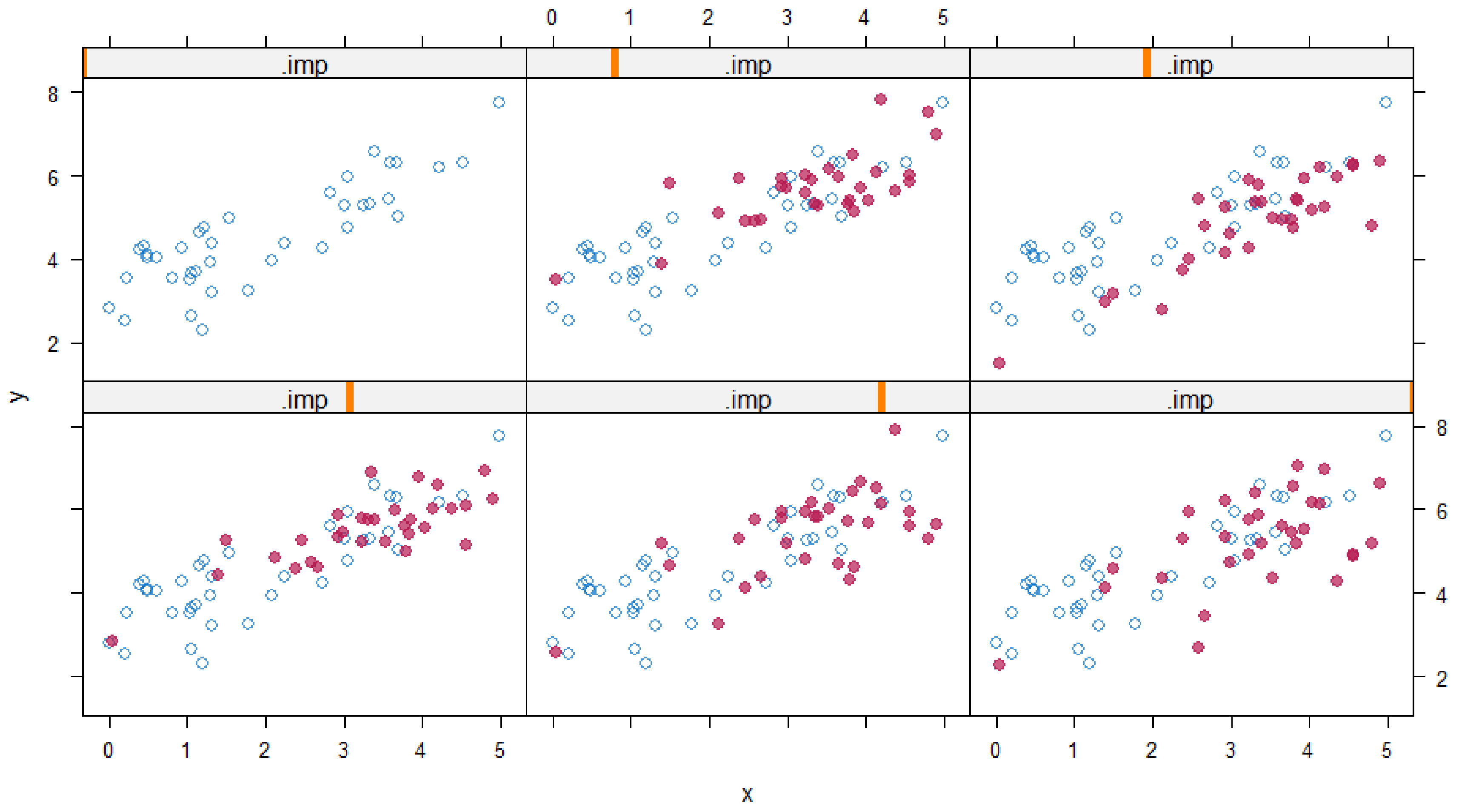
- Unbiased estimates
- CIs with correct coverage properties

There should not be any major disadvantages!

- MI is always better than single imputation

Multiple imputation (illustration)

Missing Y values: The observed data and five imputed datasets



Multiple imputation: Implementation

- Consider if the assumption of **conditional independency** of missingness is plausible
- Include in the imputation model
  - **All the variables in the analysis model**
  - Other variables (strongly) associated with the to-be-imputed variables
  - Non-linearity and interactions terms, if needed/possible
    - Especially modeling the interactions can be difficult
- The number of imputed datasets (m)
  - The more the better (although may be slow to run)
  - A rule of thumb: m = the percentage of cases with any missing values

Multiple imputation: Implementation in R

Using **mice** package (with the default settings):

Create **impute**d datasets, **fit** the models and **pool** and print the results:

```
imp <- mice(my_data, m = 20)
fits <- with(data = imp, exp = lm(y ~ x1 + x2))
summary(pool(fits))
```

- About everything can be set manually
- The default method in mice for continuous variables (Predictive mean matching) models non-linearities, non-normality and heteroskedasticy “automatically” quite well

Some guidelines/conclusion

Imputation is **not** needed or maybe not even recommended if

- Only a **small percentage** (<5%) of cases have any missing values
- Missingness **is** only in the **response** variable
  - Not much would be gained as the same model would be used for analysis and imputation
  - Missing data automatically treated by (full information) maximum likelihood (FIML)
  - **Exception:** If there are **good predictors** for the response **outside** the analysis model
- Missingness **depends** only on the **predictors**
  - Missingness is conditionally independent of the response
  - Has to be **assumed**

Some guidelines/conclusion

- Mean/median/mode imputation can be used (only) if
  - The variable is not the main predictor
  - And only small percentage missing (<5-15%??)
  - And no strong associations with other variables
- Multiple imputation is always better than single imputation
  - Appropriately done SI can be used when there are not too much missingness (<10 – 25%??, depends on the role of the variable in the analysis)
- If the assumption of conditional independence is <u>not</u> plausible
  - Even MI can/may <u>not</u> help
  - Extra assumptions ("outside the data") about the missingness need to be made and modeled

Some guidelines/conclusion

- Consider
  - How much **power** will be lost?
  - Is there any reason to be worried about **bias**?
  - Are the **assumptions** (for imputation) plausible?
- **Compare** the results with and without imputation
- How often can something useful actually be gained with imputation?

Appendix A: Statistical inference in general vs. the missing data problem

Statistical inference in general

- Does our sample (i.e data) represent the population we want? Which population does it represent?
- How much data do we need to be able to detect the effects we want?

Missing data problem

- Is the missing data from a different population? Will the estimates be biased?
- How much power do we lose if we discard the missing data?

Appendix B: Mechanisms of missingness

- Missing completely at random (**MCAR**)
  - Missingness **does not depend on any variables of interest**
  - E.g. Example 7
- Missing at random (**MAR**)
  - Missingness **depends only on the observed data,** i.e. conditional independency on the missing data
  - E.g. Examples 3, 4, 5 (if X missing) and 6 (if Y missing)
  - Required by most imputation methods
- Missing not at random (**MNAR**)
  - Missingness **depends on the missing data**
  - E.g. when missingness in a variable depends on the variable itself (Example 1, Example 5 if missingness is in Y, Example 6 if missingness is in X)
  - Imputation (without extra assumptions) usually can **not** predict missing data well

References/further reading

- van Buuren, S. (2019). Flexible Imputation of Missing Data, Second Edition. New York: Chapman and Hall/CRC
  - A very good book on multiple imputation (and missing data in general)
  - Freely available at https://stefvanbuuren.name/fimd/
- Schafer, J. L., and J. W. Graham. 2002. "Missing Data: Our View of the State of the Art." Psychological Methods 7 (2): 147–77.
  - A very good general introduction to missing data and different imputation methods
- White, I. R., and J. B. Carlin. 2010. "Bias and Efficiency of Multiple Imputation Compared with Complete-Case Analysis for Missing Covariate Values." Statistics in Medicine 29 (28): 2920–31.
  - Comparison of multiple imputation and complete case analysis