Missing data

Nan van Geloven

Overview

1. Types of missing data (recap mixed and longitudinal modelling course)

2. Methods to deal with missing data

3. Multiple imputation

4. Practical exercises:

    - missing data in DAGs

    - multiple imputation in R

Overview

1. **Types of missing data (recap mixed and longitudinal modelling course)**

2. Methods to deal with missing data

3. Multiple imputation

4. Practical exercises in R

TABLE 1.1. Would-Be Complete Data Partitioned into Observed and Missing Parts

Y_{obs} Y_{mis} M

ID	Complete			Observed			Missing			Indicators		
	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	M_1	M_2	M_3
1	13	30	15	13	30	—	—	—	15	0	0	1
2	19	38	28	19	38	28	—	—	—	0	0	0
3	20	18	8	20	18	8	—	—	—	0	0	0
4	17	39	28	—	39	—	17	—	28	1	0	1
5	22	26	12	22	26	12	—	—	—	0	0	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
496	14	36	22	—	36	22	14	—	—	1	0	0
497	28	12	7	28	—	7	—	12	—	0	1	0
498	22	30	10	22	30	10	—	—	—	0	0	0
499	24	38	13	24	38	13	—	—	—	0	0	0
500	29	8	8	—	—	8	29	8	—	1	1	0

Enders, C. K. (2022). *Applied missing data analysis.* Guilford Publications.

Missing data mechanisms

Relation between the missingness indicators M and the data $Y = (Y_{obs}, Y_{mis})$ defines the *mechanism* behind the missing data

We distinguish

      - missing completely at random (MCAR)
      - missing at random (MAR)
      - missing not at random (MNAR)

Introduced by Rubin (1976)

1. Missing Completely at Random (MCAR)

$P(M = 1 \mid Y_{obs}, Y_{mis}) = P(M = 1)$

M indicator: 0 if observed 1 if missing

Y_{obs} observed data

Y_{mis} unobserved data

- The probability of a value being missing does not depend on any other values (observed or unobserved) Y_{obs} Y_{mis}

- Observations with and without missing values are exchangeable, so those without missing values are representative for the rest

- Example: lab results gets lost in the mail

2. Missing at Random (MAR)

$P(M = 1 \mid Y_{obs}, Y_{mis}) = P(M = 1 \mid Y_{obs})$

- Chance of missing observation depends on observed data only

- Example: patients with low blood pressure are measured less frequently than patients with high blood pressure

- You can estimate what the missing values should be based on observed data (in the example, the measured blood pressure)

- Observations with and without missing values differ on certain measured characteristics

- By comparing people with and without missing values, you can distinguish MCAR and MAR (e.g. Little's test, mcartest)

3. Missing not at random (MNAR)

$M \not\perp Y_{obs}, Y_{mis}$ $M \not\perp Y_{mis}$

$P(M = 1 \mid Y_{obs}, Y_{mis})$ or $P(M = 1 \mid Y_{mis})$, do not simplify

- Missing values determine the probability of being missing even after conditioning on Y_{obs}
- Examples:
  - Depression questionnaire is not completed if patient is depressed
  - Doctor does not record weight of a patient, because he sees that the patient is not obese

- Difficult. You cannot reconstruct the missing values using measured factors, have to make assumptions that you cannot verify
- Perform sensitivity analysis to calculate results under different assumptions for the missing mechanism
- Cannot determine from the data whether missing at random or not at random (!)

MCAR, MAR, MNAR?

- Some depressed people have not filled in their income in a survey

  MAR

- Depressed people with a low income are less likely to fill in their income

  MNAR

  $$P(M=1 \mid Y_{mis})$$

- An MRI measurement is performed in 30% of the participants in a study

  MCAR

Overview

Which target parameters can still be estimated consistently from observed data?

Depends on

1. missings in outcome or exposure or confounder,

2. missings dependent on outcome? *serious*

3. missing mechanism (MCAR, MAR, MNAR)

4. the target parameter itself
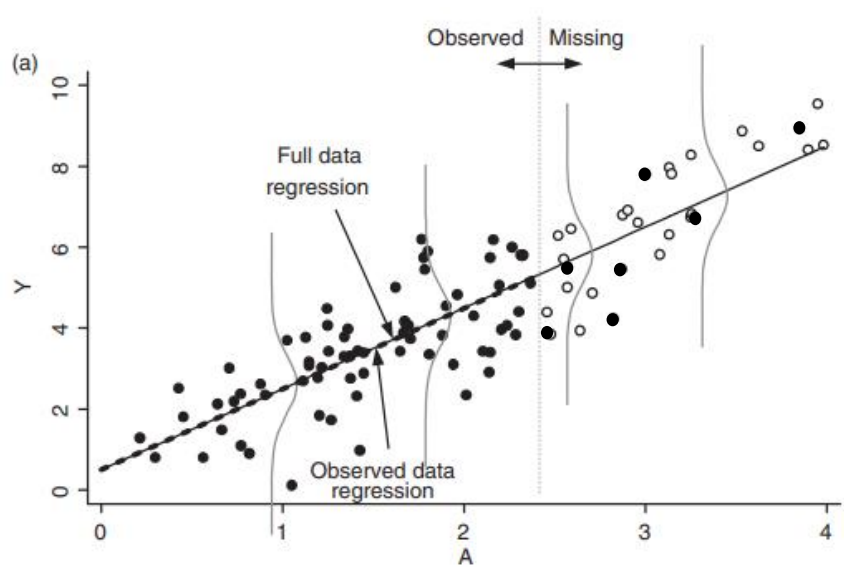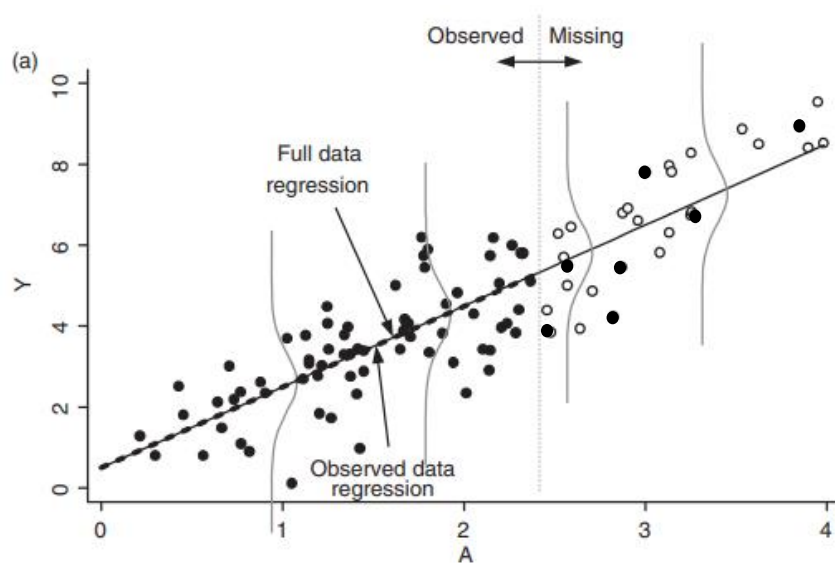
Y – continuous outcome , A – continuous exposure

Example 1:

Y values are missing for 75% of observations with A>2.5

　Missing mechanism? *MAR*

　Can we estimate the mean of Y in full population from observed data? *X*

　Can we estimate the effect of A on Y from observed data?

In the case, ✓

Examples taken from: Rhian et al. Using causal diagrams to guide analysis in missing data problems. SMMR 2011

Y – continuous outcome , A – continuous exposure

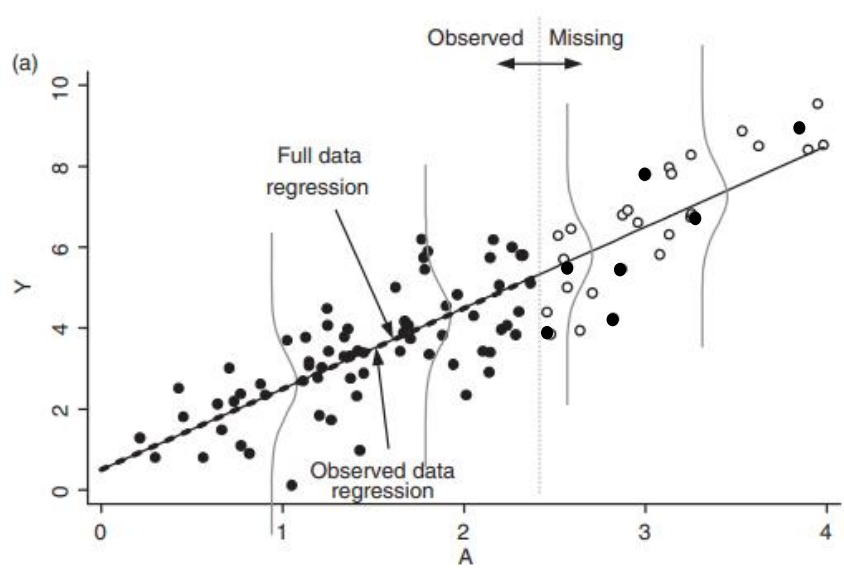Example 1:

Y values are missing for 75% of observations with A>2.5

Missing mechanism? -> MAR

Can we estimate the mean of Y in full population from observed data? -> no

Can we estimate the effect of A on Y from observed data? -> yes

Y – continuous outcome , A – continuous exposure

Example 2:

A values are missing for 75% of observations with A>2,5

Missing mechanism? *MNAR*

Can we estimate the mean of Y in full population from observed data? ✓

Can we estimate the effect of A on Y from observed data? ✓

trend is stable

Y – continuous outcome , A – continuous exposure

Example 2:

A values are missing for 75% of observations with A>2,5

    Missing mechanism? -> MNAR

    Can we estimate the mean of Y in full population from observed data? -> yes

    Can we estimate the effect of A on Y from observed data? -> yes

Y – continuous outcome , A – continuous exposure

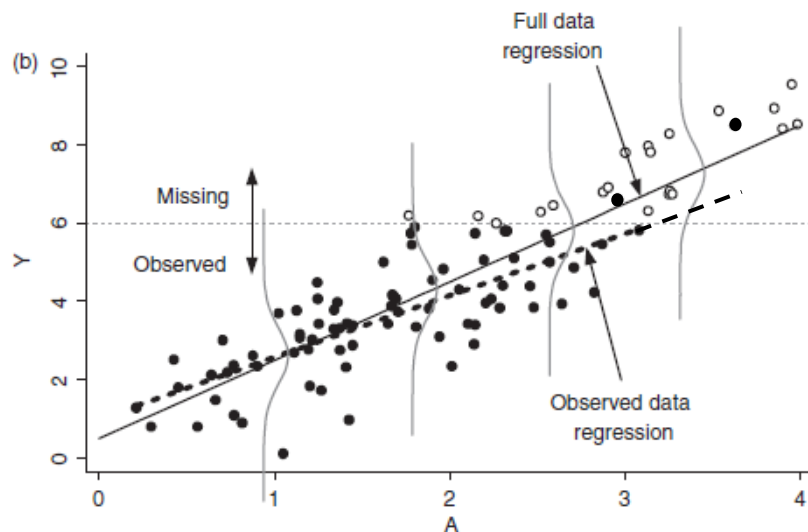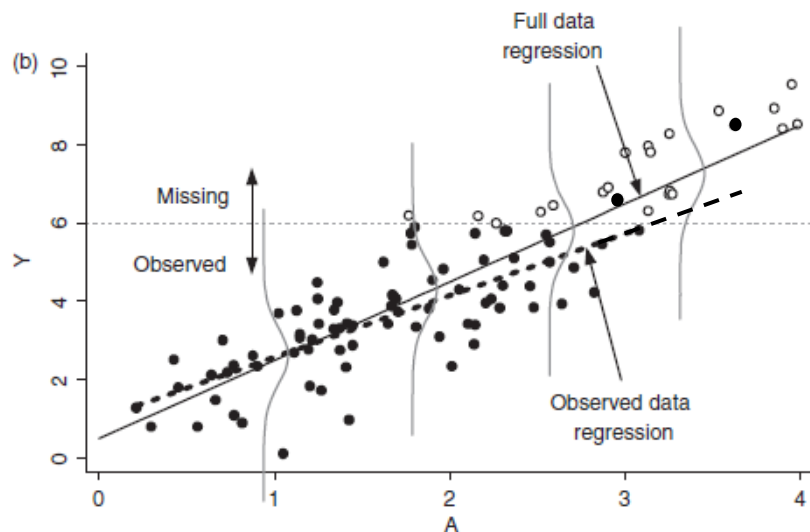Example 3:

90% of Y values are missing for Y>6

Missing mechanism? MNAR

Can we estimate the mean of Y in full population from observed data? X

Can we estimate the effect of A on Y from observed data? X

(b)

Full data regression

Missing

Observed

Observed data regression

Y

A

Y – continuous outcome , A – continuous exposure

Example 3:

90% of Y values are missing for Y>6

Missing mechanism? -> MNAR

Can we estimate the mean of Y in full population from observed data? -> no

Can we estimate the effect of A on Y from observed data? -> no

How to handle missing values in the analysis?

Use only observed data / "complete cases"

- If all incomplete variables are expected to be MCAR

- will decrease efficiency (ie larger standard errors / p-values etc)

- If the number of missing observations is small (<5%), bias (and loss of efficiency) will not be large

YES, if possible

How to handle missing values in the analysis?

- Make a separate category/ indicator variable for missing

    - Used in some prediction algorithms[1]

    - Interpretation typically only useful if the missing mechanism in future data is similar

- Fill in average (or median) values for missing data

    - May lead to dilution of effects if used on exposure variable

    - Useful in certain specific cases, e.g. missing values in covariates that only affect the outcome[2]

    - Single imputation will yield too small estimates of standard errors for imputed variables
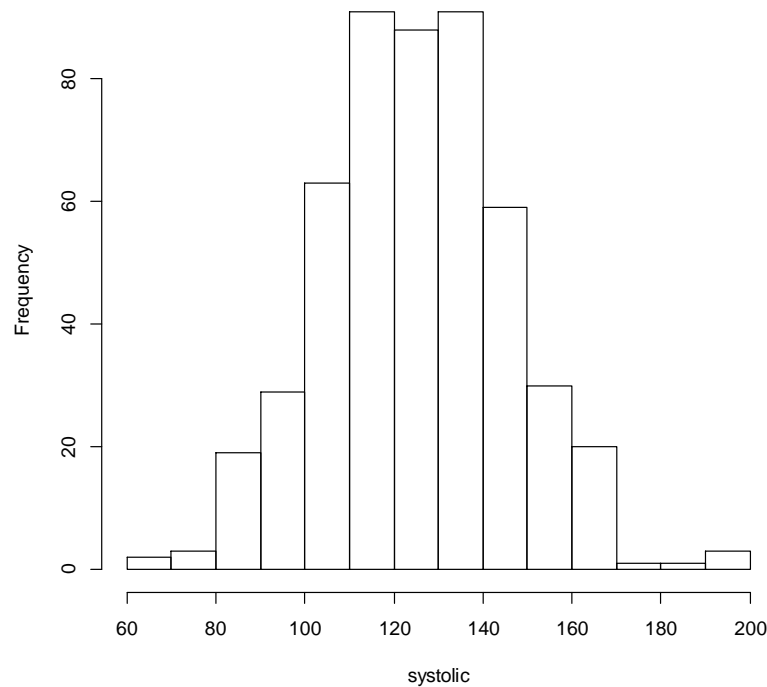
Only in special cases

1. Twala et al 2008 Pattern Recognition letters. Good methods for coping with missing data in decision trees
2. White ad Thompson 2004 Stat Med Adjusting for partially missing baseline measurements in randomized trials

Single imputation will yield too small estimates of standard errors
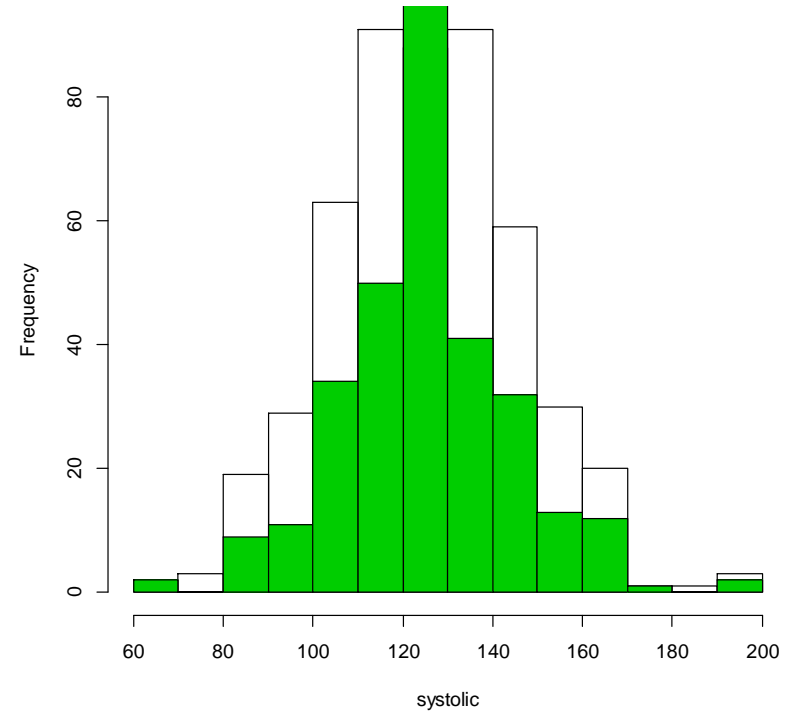
replace half of the observations by the mean

mean=125 sd=21

mean=125 sd=16

Other approaches (assuming MAR)

- Maximum likelihood methods:
    - Mixed models (for missing values in repeatedly measured outcome variables)
    - Difficult to use when missing values occur in several variables (exposures, confounders and outcome)

- Inverse probability weighting
    - Estimate for each observation the probability to have missing value
    - Observations with a high probability to be missing will get higher weight ? ?

- Multiple Imputation

Yes, generally safe

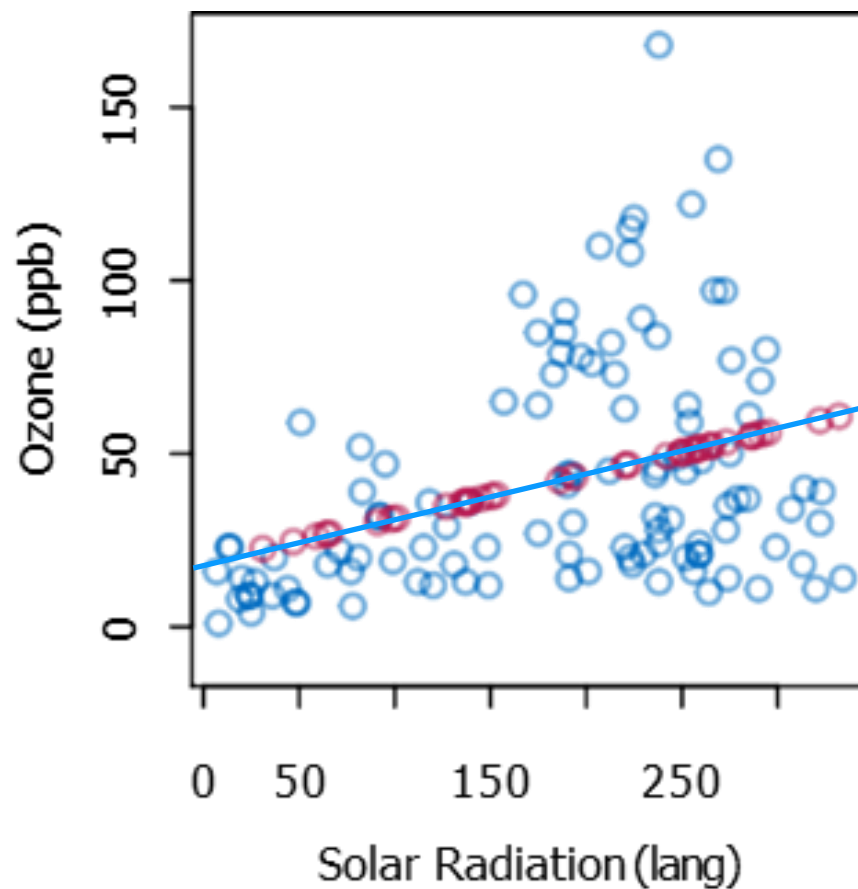# In summary: when to use which analysis method

- First: think about the most plausible mechanism that led to the missing observations

- Choose appropriate analysis strategy

  - MCAR in all variables -> complete cases are unbiased, consider multiple imputation only for efficiency gain

  - MAR/MNAR but not dependent on outcome -> complete cases may still be unbiased for some analyses, consider multiple imputation for efficiency gain

  - MAR with dependency on outcome:

    - maximum likelihood method available?

    - no -> multiple imputation

  - MNAR with dependency on outcome:

    - sensitivity analysis

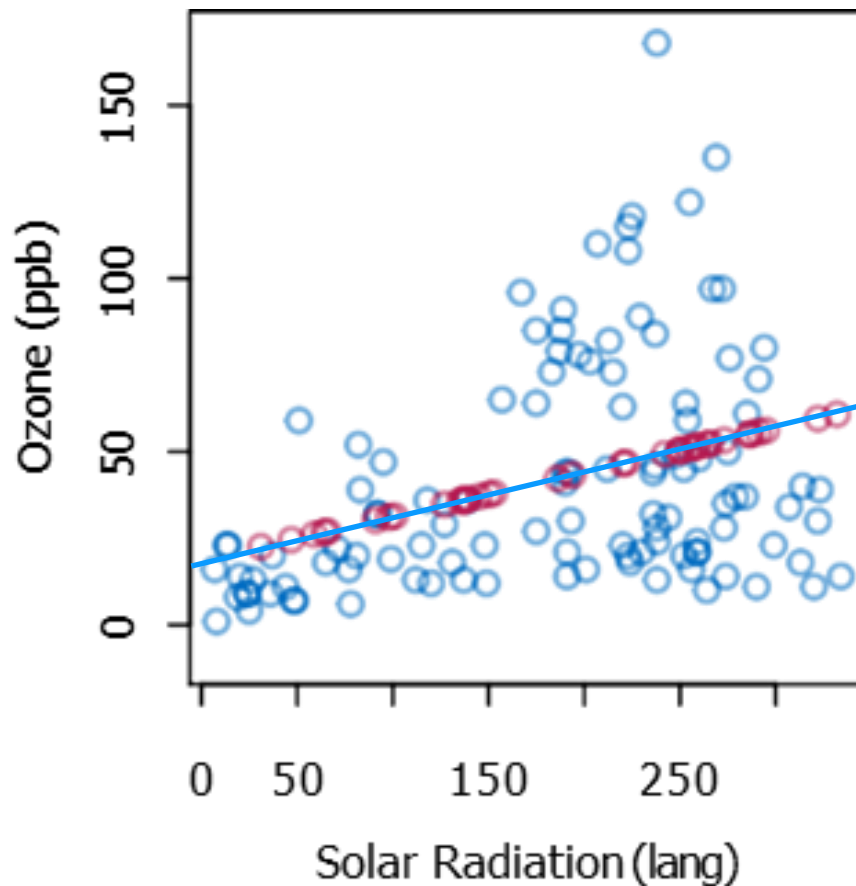 Rule of thumb: <5% missing data -> complete cases probably ok in any case

Overview

1. Types of missing data (recap repeated measurements course)

2. Methods to deal with missing data

3. **Multiple imputation**
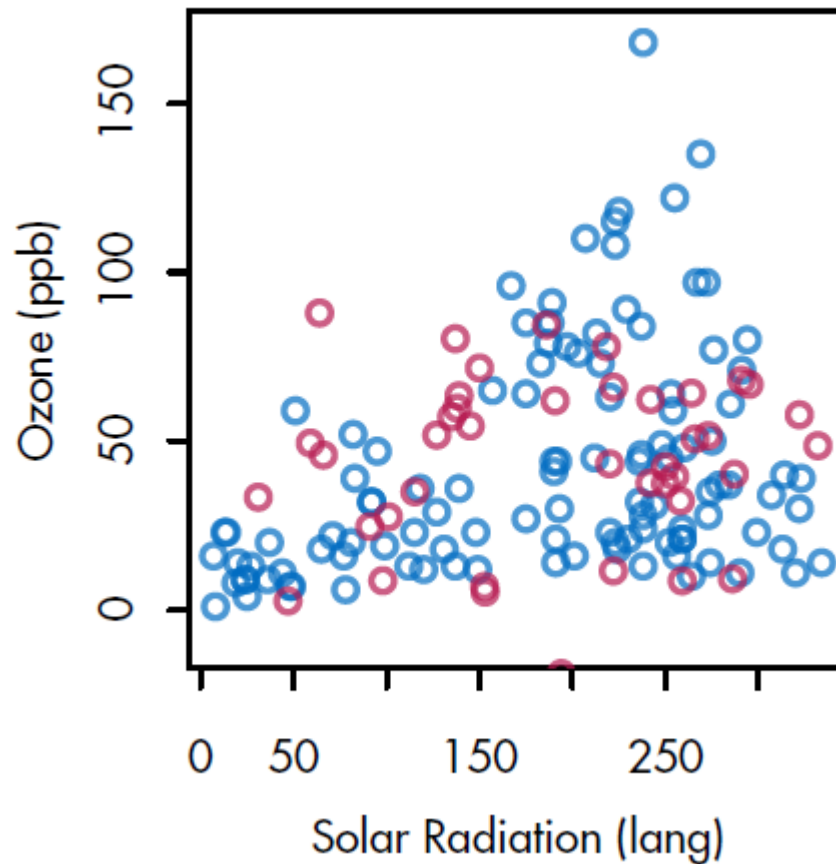
4. Practical exercises in R

What if we fill in missing values based on regression between exposure and outcome?

- means ok
- too strong correlation
- wrong confidence intervals / p-values

- means ok
- correlation ok
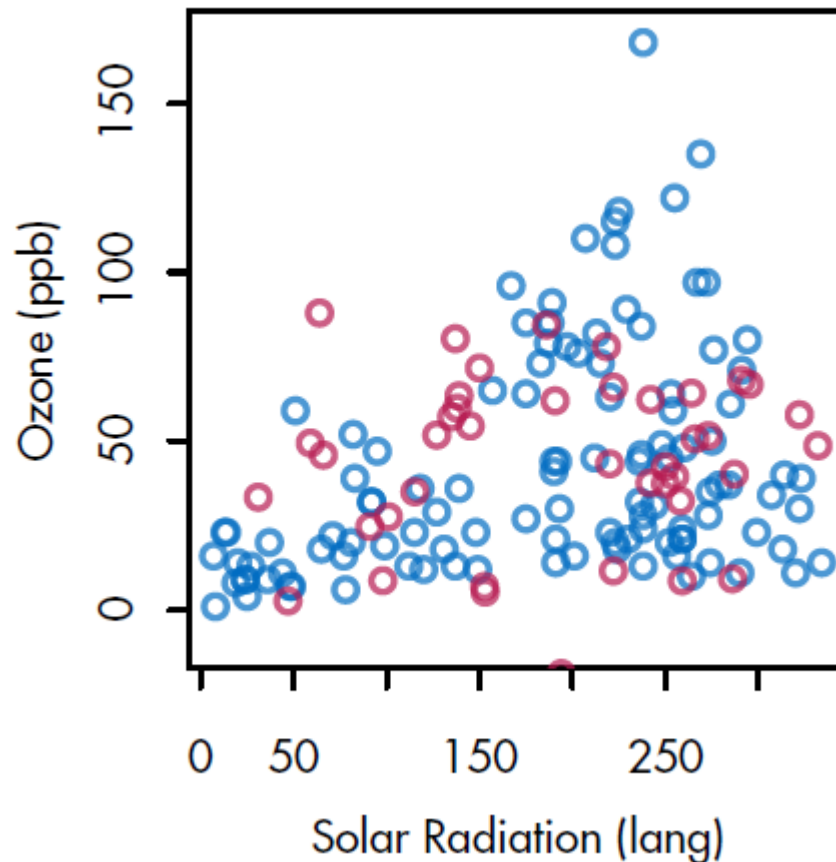
- means ok
- correlation ok
- BUT: still wrong confidence intervals / p-values!!

- Need to 'confess' that we did not observe all these data!
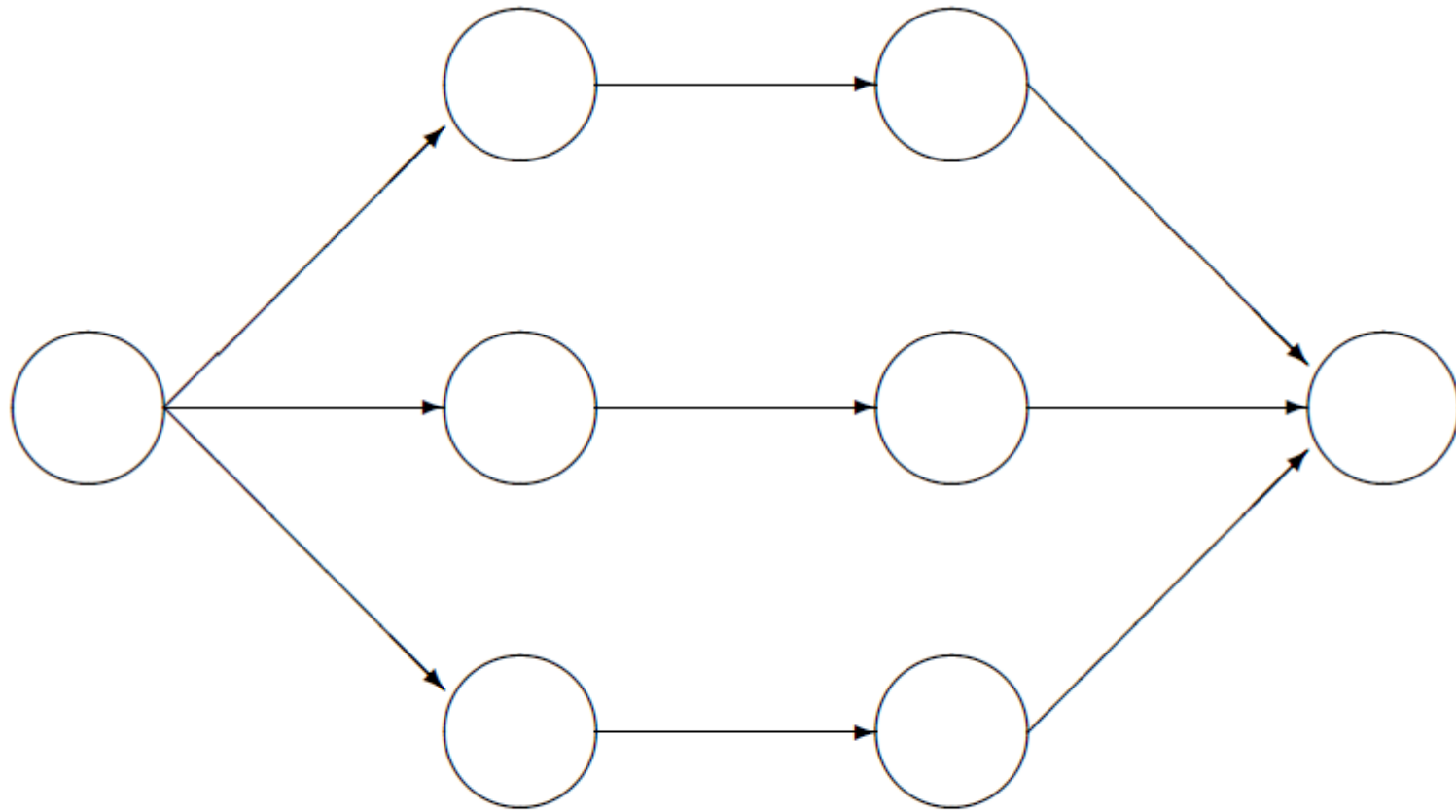
Incomplete data   Imputed data   Analysis results   Pooled results

Incomplete data   Imputed data   Analysis results   Pooled results

Incomplete data   Imputed data   Analysis results   Pooled results

Incomplete data   Imputed data   Analysis results   Pooled results

Incomplete data   Imputed data   Analysis results   Pooled results
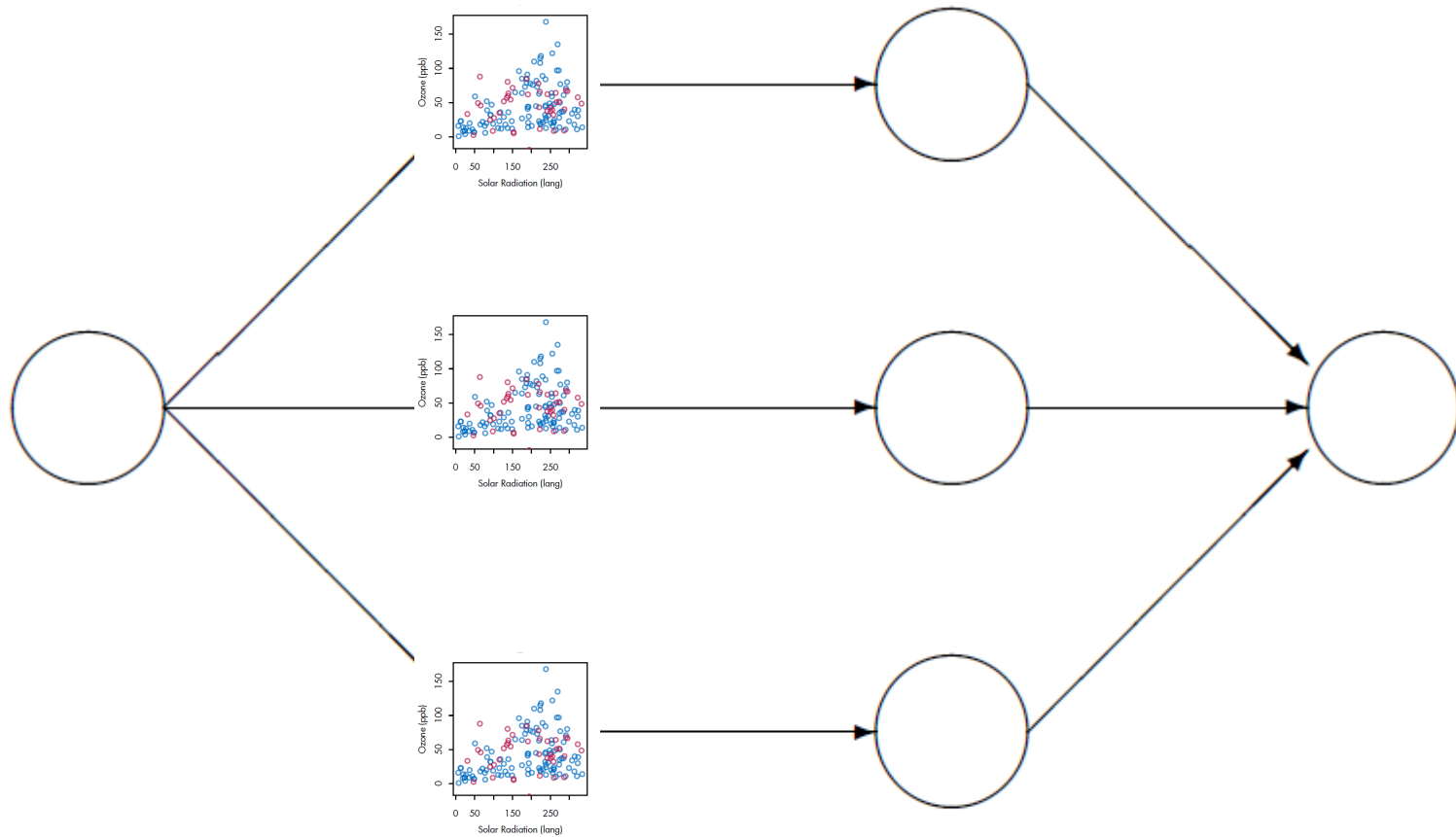
How is the missing Y value being imputed? Regression method:

1. Formulate imputation model

    * $y = \beta_0 + \beta_1 x + e$

    * $e \sim N(0, \tau^2)$

2. Fit the model, using observed data.

3. Draw values for β_0, β_1, and τ^2, based on estimated coefficients and standard errors

4. Use the resulting imputation model to draw values for missing observations

5. This yields a complete dataset

6. Analyse the complete dataset and obtain an estimate with standard error

Example: Y is missing for x=5
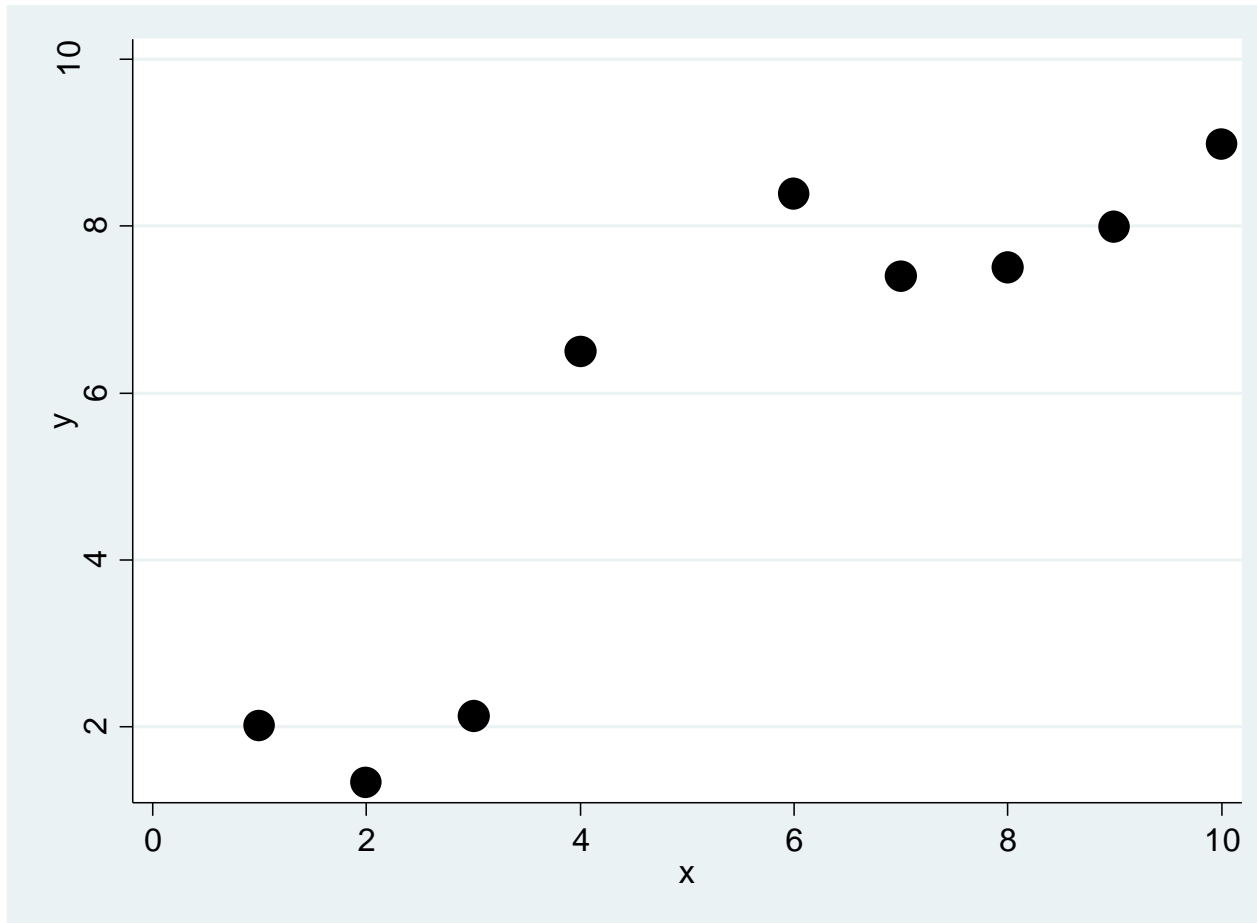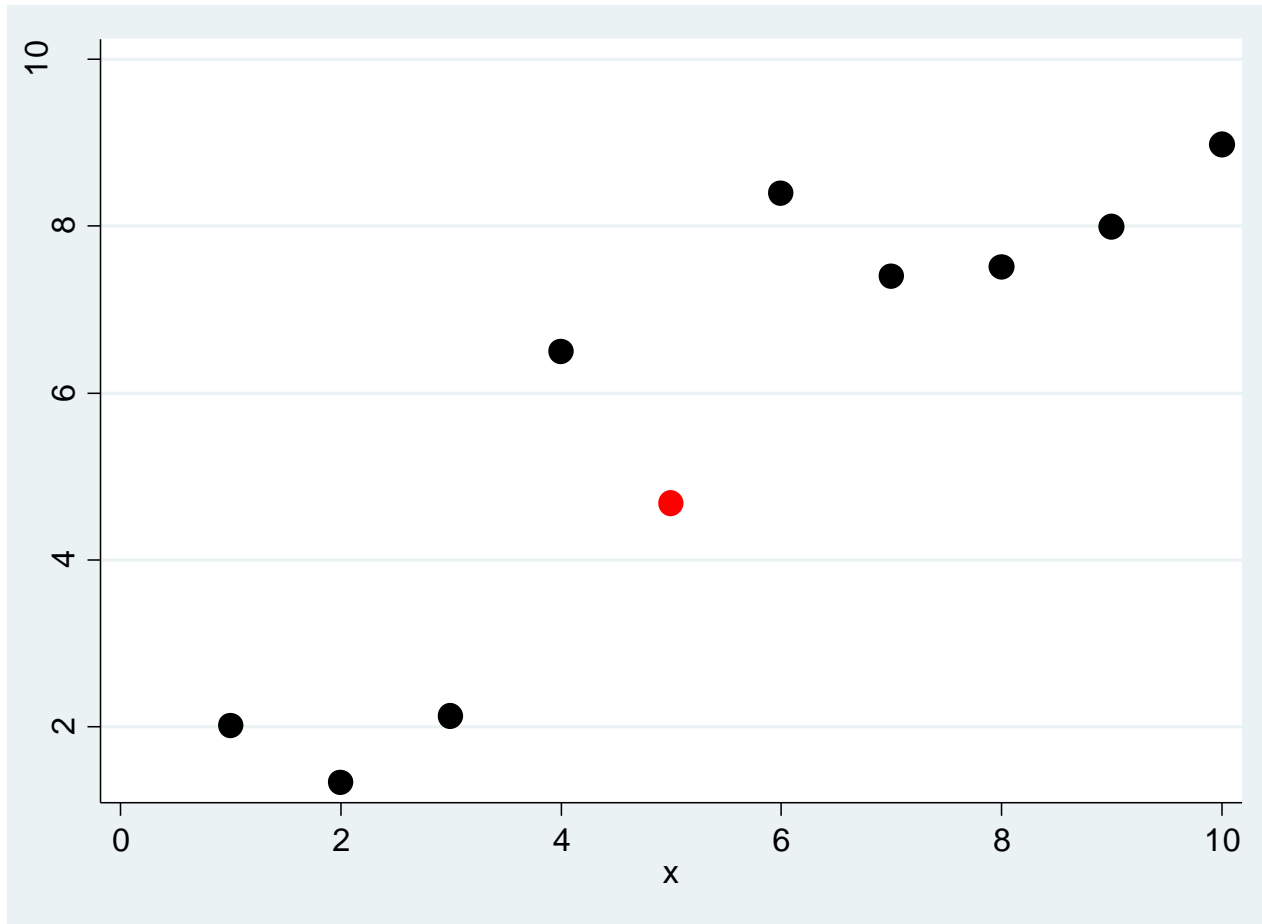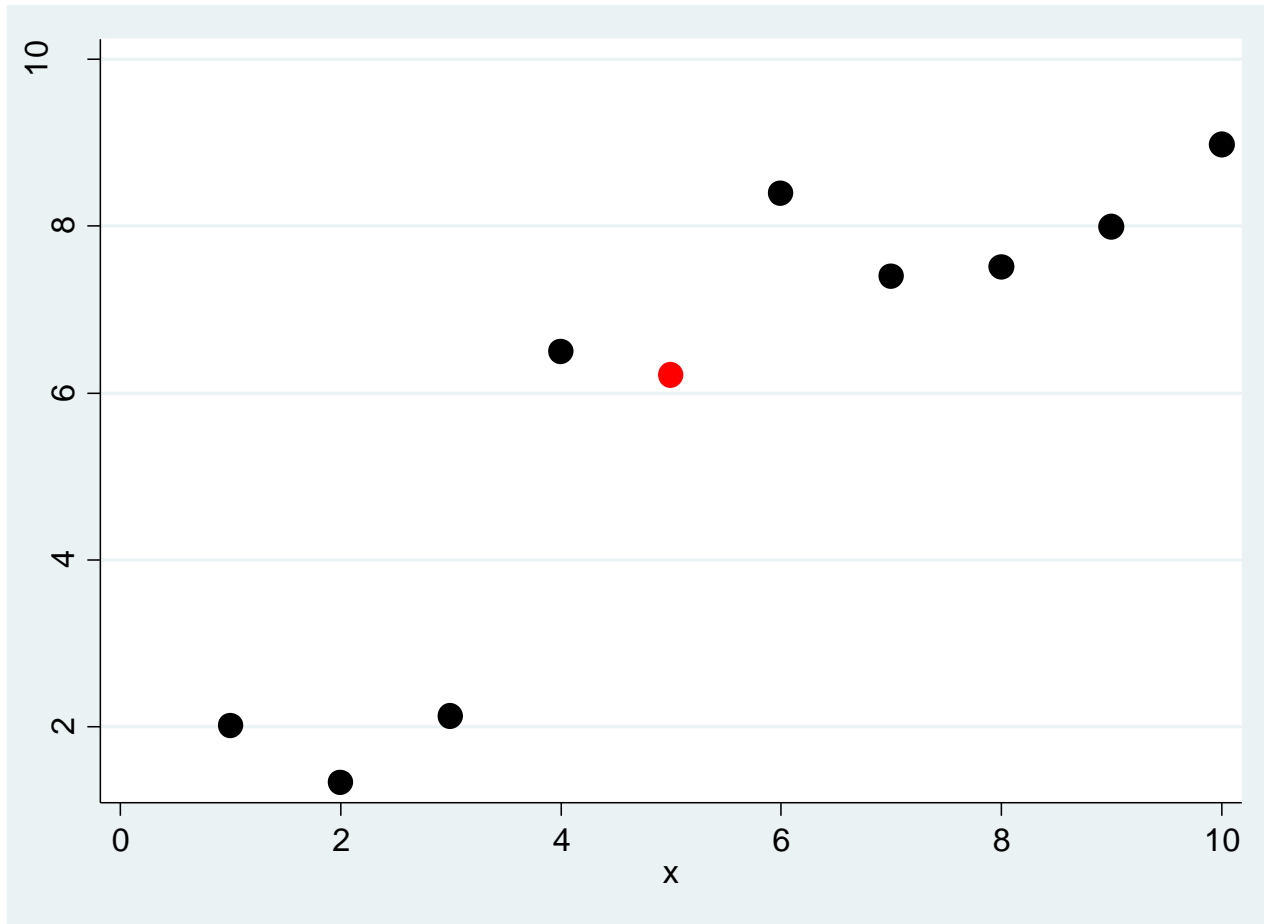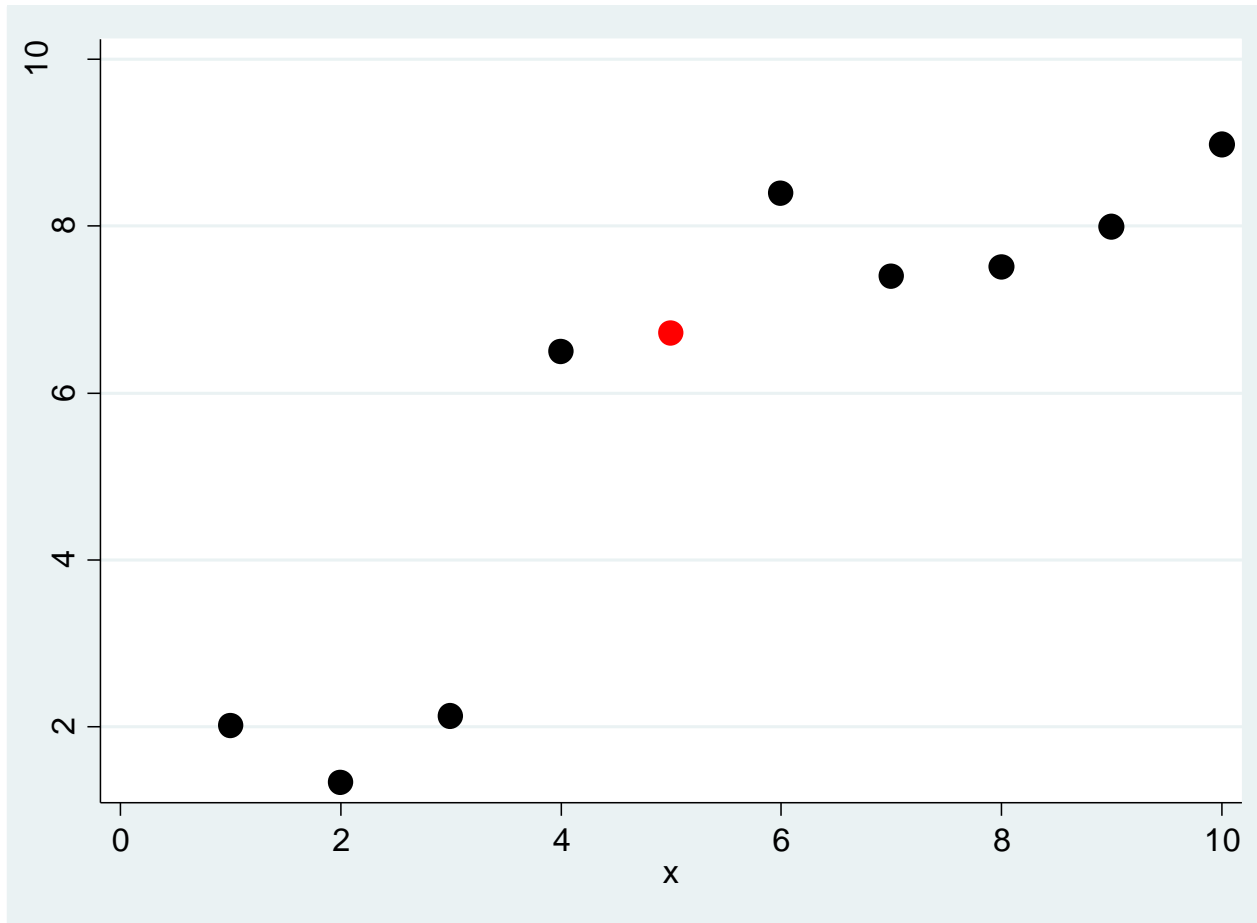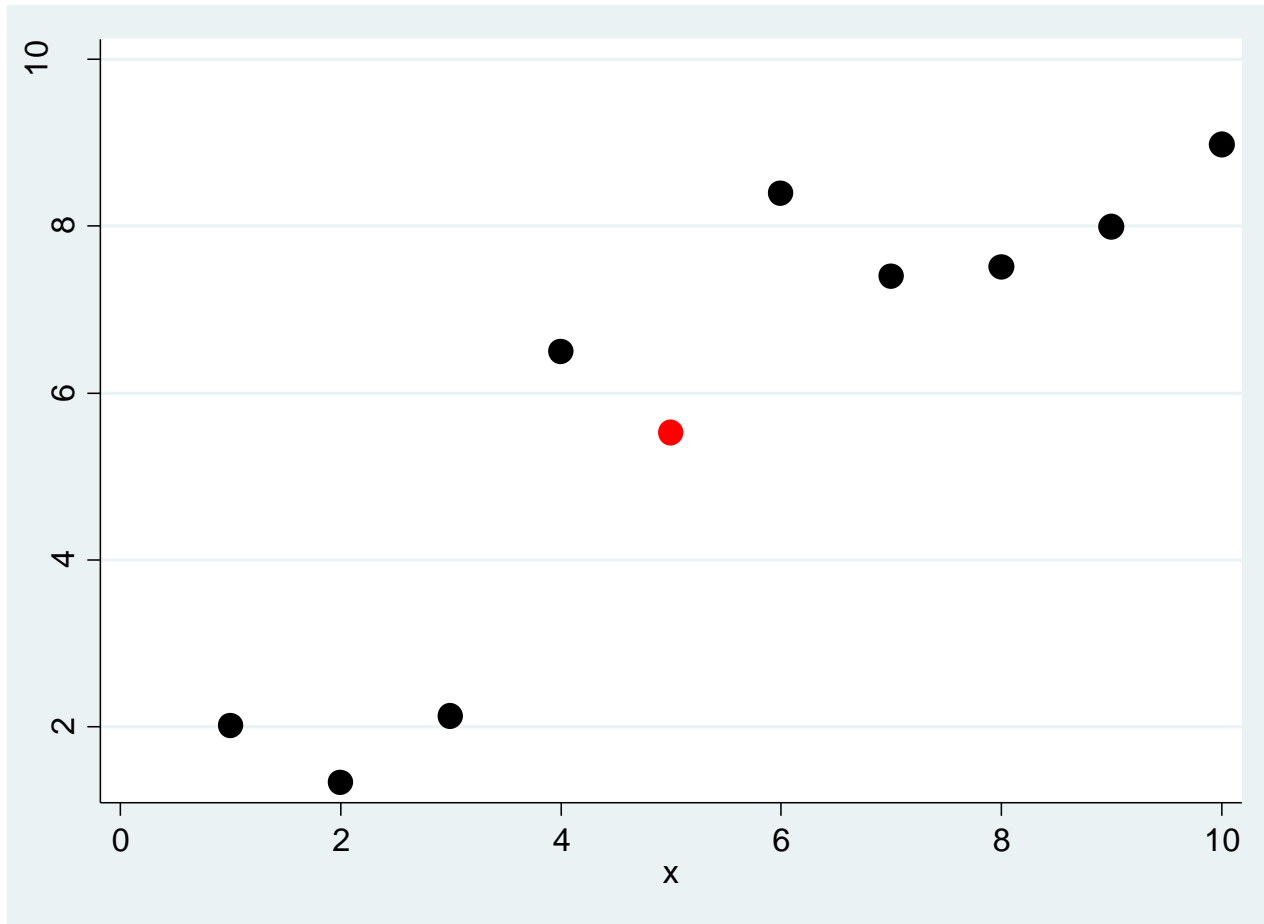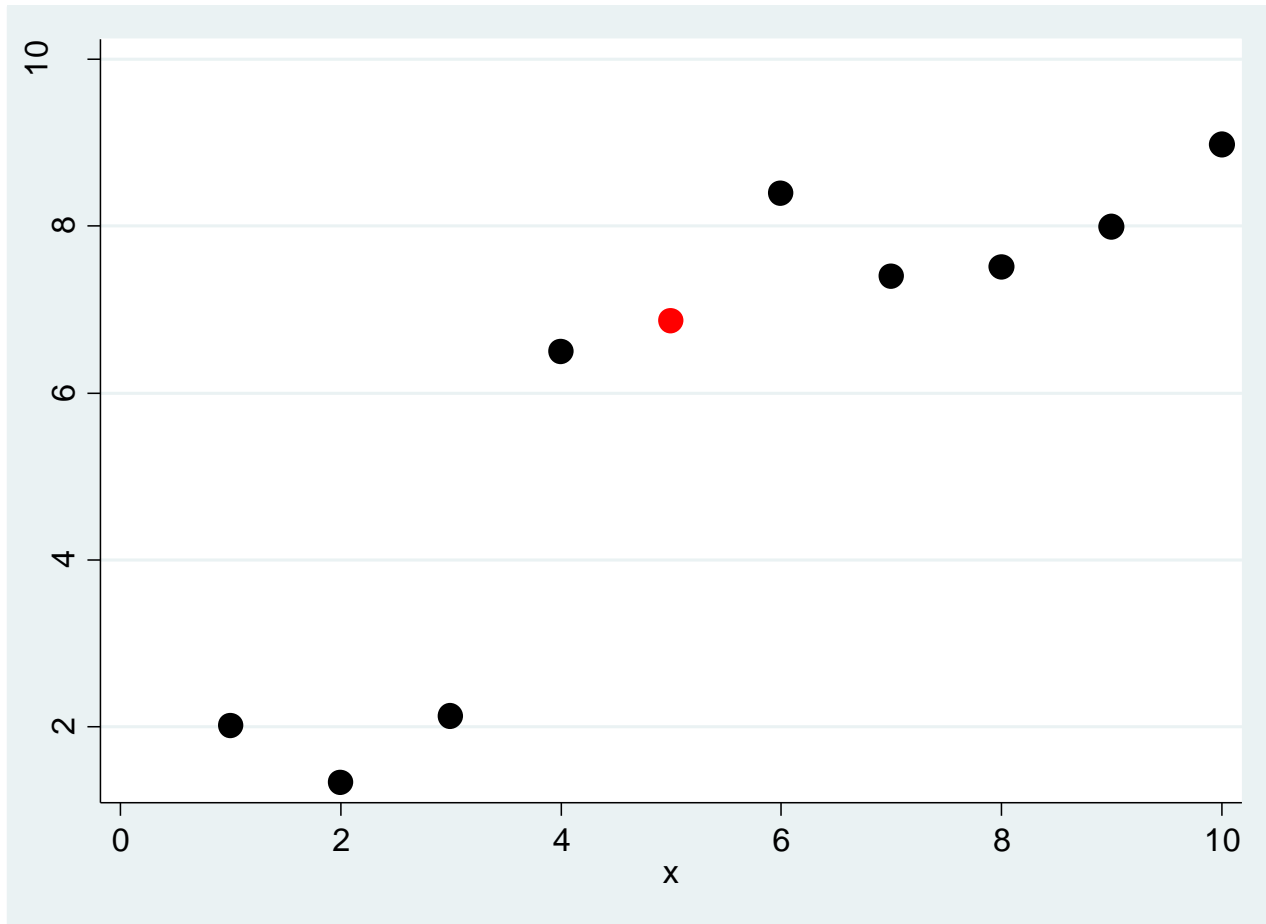
Rubin's rules

For each dataset $i = 1, \ldots, m$ we have a β_i and se_i

Now want to pool these:

Pooled β simply the average:

$$\beta = \frac{1}{m} \sum_{i=1}^{m} \beta_i$$

Pooled se slightly more complicated:

$$se(\beta) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} se_i{}^2 + \left(1 + \frac{1}{m}\right) sd^2}$$

with sd the standard deviation of $(\beta_1, \beta_2, \ldots \beta_m)$

$\beta_1 + se_1$

$\beta_2 + se_2$ → $\beta + se$

$\beta_3 + se_3$
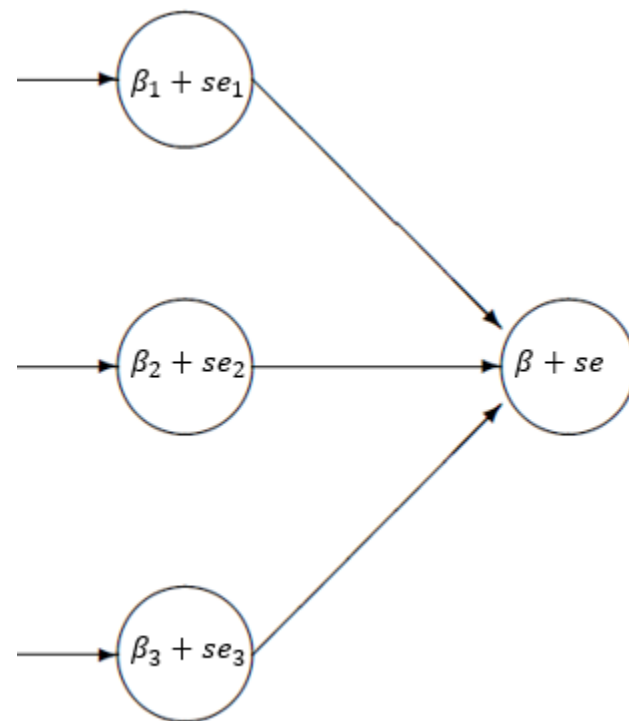
Rubin's rules example

example

	β	se
Imputation 1	5.69	.92
Imputation 2	5.85	.91
Imputation 3	5.90	.92
Imputation 4	5.78	.91
Imputation 5	5.91	.92

Average of the β's is 5.83 (sd=0.09)
Average of the se^2 is 0.916^2

- Pooled estimated β is 5.83

- se of the pooled β is : $\sqrt{0.916^2+(1+1/5)0.09^2}=1.01$

What if there are several variables with missing values?

- Approach 1: Impute, assume multivariate normal distribution
  - Will also work if variables are ordinal or binary
  - For skewed distributions: first transformation

Monotone imputation method

Monotone Missing Data Patterns

Group	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	.

- Impute first Y_2 with Y_1, then Y_3 with Y_2 and Y_1, etc.

- Be aware: the order of the variables matters

Chained equations (MICE)

- Fill in random values for all missing data

- Impute missing data for X_1 with $X_2 \ldots X_n$

- Impute missing data for X_2 with imputed X_1 and $X_3 \ldots X_n$ etc.

- Cycling through these steps until the process is stable

- Procedure known *multivariate imputation by chained equations* (MICE) or *fully conditional specification* (FCS)

- NB: transform skewed variables first, and transform back after imputation

Y	X_1	X_2	X_3
-0,3	0	0,8	-0,2
-1,9	1	-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1	-0,8	1,2
-0,3	1	-1,4	1

Y	X_1	X_2	X_3
-0,3		0,8	-0,2
-1,9		-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1		1,2
-0,3	1		1

Y	X_1	X_2	X_3
-0,3	0	0,8	-0,2
-1,9	0	-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1	0,4	1,2
-0,3	1	0,8	1

Initialize by sampling from the observed values

Y	X₁	X₂	X₃
-0,3	0	0,8	-0,2
-1,9	0	-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1	0,4	1,2
-0,3	1	0,8	1

Model $X_1 \mid Y, X_2, X_3$ (e.g. a logistic regression) based only on observed X_1

Y	X₁	X₂	X₃
	Y	X_1	X_2	X_3

Y	X_1	X_2	X_3
-0,3	1	0,8	-0,2
-1,9	0	-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1	0,4	1,2
-0,3	1	0,8	1

Impute! (accounting for parameter uncertainty/random variation)

Y	X_1	X_2	X_3
-0,3	1	0,8	-0,2
-1,9	0	-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1	0,4	1,2
-0,3	1	0,8	1

Model $X_2 \mid Y, X_1, X_3$ based using observed X_2 (e.g. a linear regression) - note that the latest imputed X_1 are used in the model

Y	X_1	X_2	X_3
-0,3	1	0,8	-0,2
-1,9	0	-1,1	-1,4
1,8	1	0,4	0,4
1,4	1	-0,3	0,6
0,3	0	0,4	1,5
-0,1	0	0,4	0,6
-1,8	0	-0,5	0,9
-0,4	0	-0,9	0,8
0,3	1	0,5	1,2
-0,3	1	0,6	1

Impute! This is now a single imputed dataset based on one iteration/cycle across the variables

General rule: enter in imputation model:

- Variables that will be used in analysis model (including the outcome !!!)

- Variables related to the missing process

- Other variables associated with the variables with missing values

- Complexity of the analysis model must also be included in the imputation (e.g. interaction terms or random effects)

Which model to use to impute?

- Continuous variables: e.g. linear regression
- Binary variables: e.g. logistic regression
- Categorical variables: e.g. multinomial logistic regression or ordinal logistic regression
- Counts: e.g. Poisson regression

- More flexible algorithms can also be used, for example *mice* package in R supports trees, forests, LASSO etc.

Predictive mean matching

- Alternative to imputing numeric values with a regression model: predictive mean matching (PMM)

- Uses the drawn predicted y for a missing value
- Then imputes the observed y of the person closest in prediction to the imputed value (or the mean of several persons closest by)

- Advantage: you will not obtain unrealistic values of y
- Disadvantage: variability may be underestimated

- PMM is default method to impute numeric values in the *mice* package in R

Some final tips

- Number of imputed datasets?

    - Rubin: 5

    - Use for datasets with many variables with missing values more imputation sets (50 or more)

    - If feasible, you may vary the number of imputations to see if this affects your results

    - Very large number helps for reproducibility but not per se for accuracy

- Imputation model should be well specified

    - Check the underlying models, are they stable?

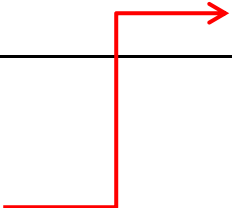   - Useful to first run one model for each variable separately

Some cautions

- More complicated settings:

  - survival data

  - clustered/hierarchial data

  - combining MI with bootstrapping , making predictions/validation

  - For propensity scores (Choi, Dekkers, le Cessie, 2018, Eur J Epidemiol)

Caveat: much can be pooled, but not everything

Table VIII. Common statistics that can and cannot be combined using Rubin's rules (equations (1) and (2)).

Statistics that can be combined without any transformation	Mean, proportion, regression coefficient, linear predictor, C-index, area under the ROC curve
Statistics that may require sensible transformation before combination	Odds ratio, hazard ratio, baseline hazard, survival probability, standard deviation, correlation, proportion of variance explained, skewness, kurtosis
Statistics that cannot be combined	P-value, likelihood ratio test statistic, model chi-squared statistic, goodness-of-fit test statistic

"Things that change when sample size increases"

Solution: present range over imputation runs or present results for one random set (e.g. for presenting a figure)

White, Royston, Wood 2011

In summary

- First: think about the most plausible mechanism that led to the missing observations

- Second: make the DAG (explained in practical exercises)

- Choose appropriate analysis strategy (see slide 22)

- Compare observed data with imputed data: understand any differences

- Report on the assumptions made

Reading material

Reading material for exam + helpful for practical exercise:

- [Lee KJ et al. The Treatment And Reporting of Missing data in Observational Studies framework. JCE 2021](#)

Hands-on tips for performing multiple imputation:

- [White IR et al. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011](#)

Other sources

- A gentle introduction to imputation of missing values.
  Donders et al. JCE 2006
- Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Sterne et al BMJ 2009;338:b2393
- When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. Jakobsen et al BMC Med Res Methodol. 2017.

Overview

After practical exercises: what did we learn?

- There are situations where complete case analysis yields unbiased results, even for MAR or MNAR

- Imputation will give the correct estimates(unbiased) for MAR and MCAR **if the imputation model is correct**

- Multiple imputation is not working well for MNAR (imputation methods to handle MNAR exist, but they require extra untestable assumptions)

- Important to think carefully about which variables to use in the imputation model, otherwise we are jumping out of the fryer pan into the fire (van de regen in de drup)

- The outcome should be added to the imputation model.