MA8701 Advanced methods in statistical inference and learning

Week 3 ~~(L5+L6)~~: Missing data

L5

Mette Langaas

~~1/22/23~~ 23.01.2023

Course homepage:
https://wiki.math.ntnu.no/ma8701/2023v/start

Reading list:

L6

- Handbook of missing data methodology: Chapter 12.1.2, 12.2, 12.3.3. Available for download (60 pages pr day) from Oria at NTNU (Choose EBSCOweb, then PDF full text and then just download chapter 12).

Stef van Buuren L5

- Flexible imputation of missing data"): Chapters 1.1, 1.2, 1.3, 1.4, 2.2.4, 2.3.2 (similar to Handbook 12.2), 3.2.1, 3.2.2 (Algo 3.1), 3.4 (Algo 3.3), 4.5.1, 4.5.2.

L6

Missing data

Many statistical analysis methods (for example regression) require the data (for analysis) to be *complete*. That is, for all data record (observations, rows) all the variable under study must be *observed*. But, in the real world this is not the case - some variables are missing for some observations (records, rows).

What are reasons for data to be missing?

Some reasons (not exhaustive):

- nonresponse,
- measurement error,
- data entry errors,
- data collection limitations,
- sensitive or private information,
- data cleaning.

The missingness may be intentional (sampling) or unintentional (refusal, self-selection, skip questions, coding error).

What can we do if we have missing data?

Use a method that handles missing data (why missing)

CART, xgboost

Fill in values to create ~~a~~ one full data set

single imputation

- LOCF
- mean
- regression

Delete all incomplete records

complete case analysis

Fill in values to create m full data sets

multiple imputation

Indicator variable method

Fix the likelihood

Fully Bayes methods

How will this affect the analysis?

- Complete case: wrong conclusions
  loss of power

- Single imputation: too confident

- Multiple imputation: complicated analysis

Airquality

A data frame with 153 observations on 6 variables.

`` `[,1]` ``	`` `Ozone` ``	numeric	Ozone (ppb)
`` `[,2]` ``	`` `Solar.R` ``	numeric	Solar R (lang)
`` `[,3]` ``	`` `Wind` ``	numeric	Wind (mph)
`` `[,4]` ``	`` `Temp` ``	numeric	Temperature (degrees F)
`` `[,5]` ``	`` `Month` ``	numeric	Month (1--12)

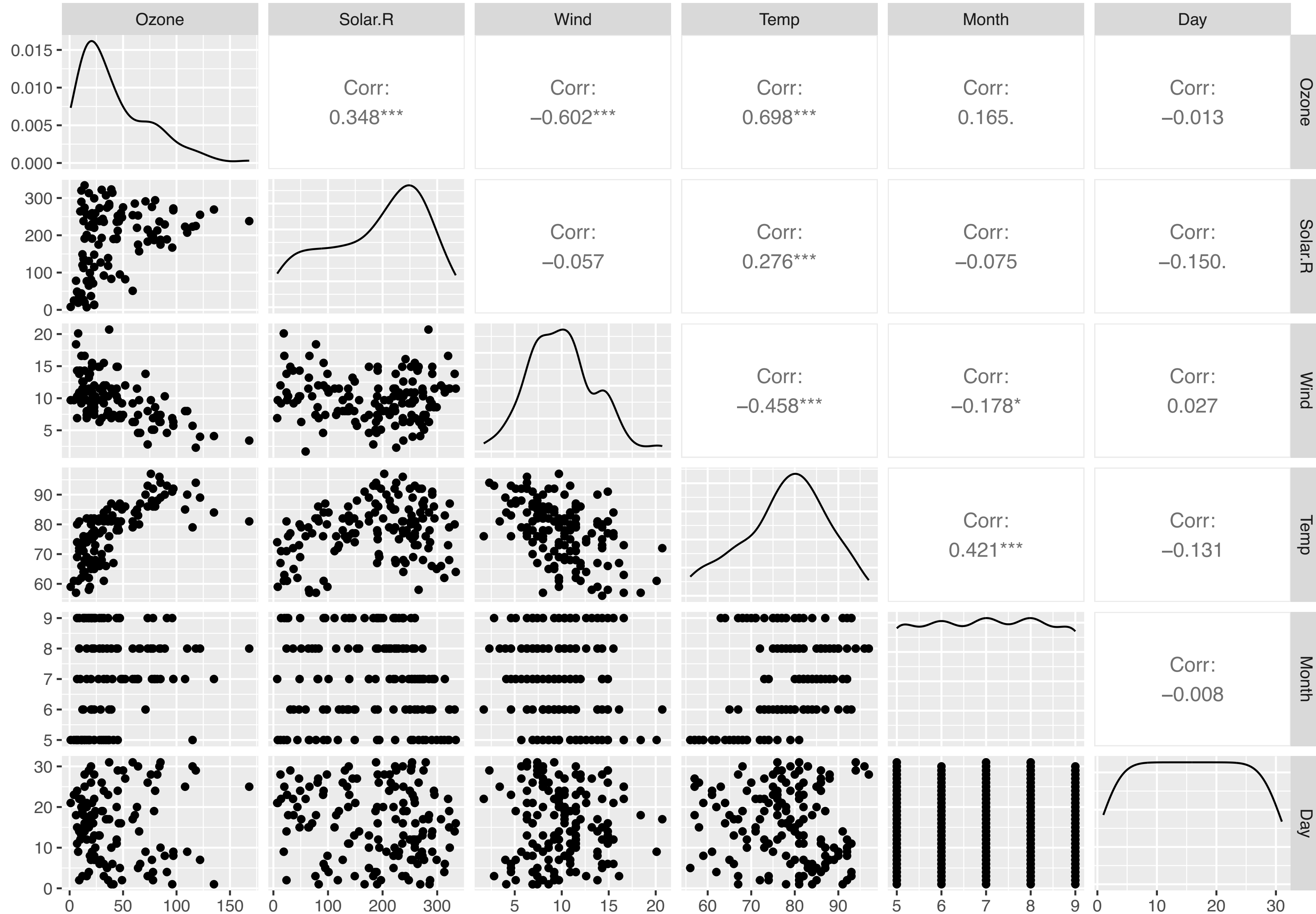
Details

Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.

- `` `Ozone` ``: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- `` `Solar.R` ``: Solar radiation in Langleys in the frequency band 4000--7700 Angstroms from 0800 to 1200 hours at Central Park
- `` `Wind` ``: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
- `` `Temp` ``: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

References

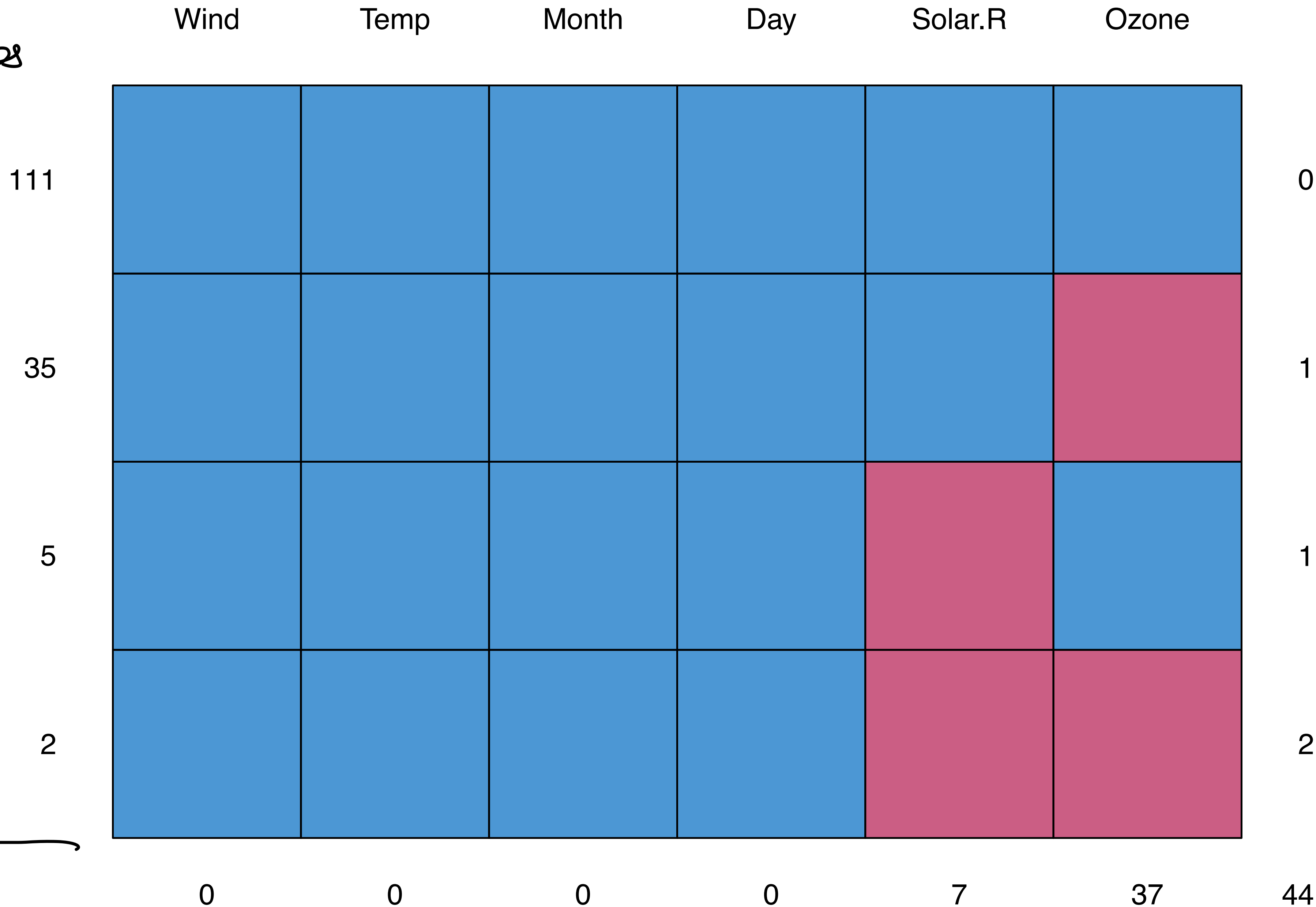
Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.



Ozone
Solar.R
Wind
Temp
Month
Day
Corr:
0.348***
Corr:
−0.602***
Corr:
0.698***
Corr:
0.165.
Corr:
−0.013
Corr:
−0.057
Corr:
0.276***
Corr:
−0.075
Corr:
−0.150.
Corr:
−0.458***
Corr:
−0.178*
Corr:
0.027
Corr:
0.421***
Corr:
−0.131
Corr:
−0.008
Ozone
Solar.R
Wind
Temp
Month
Day

MISSING PATTERN PLOT

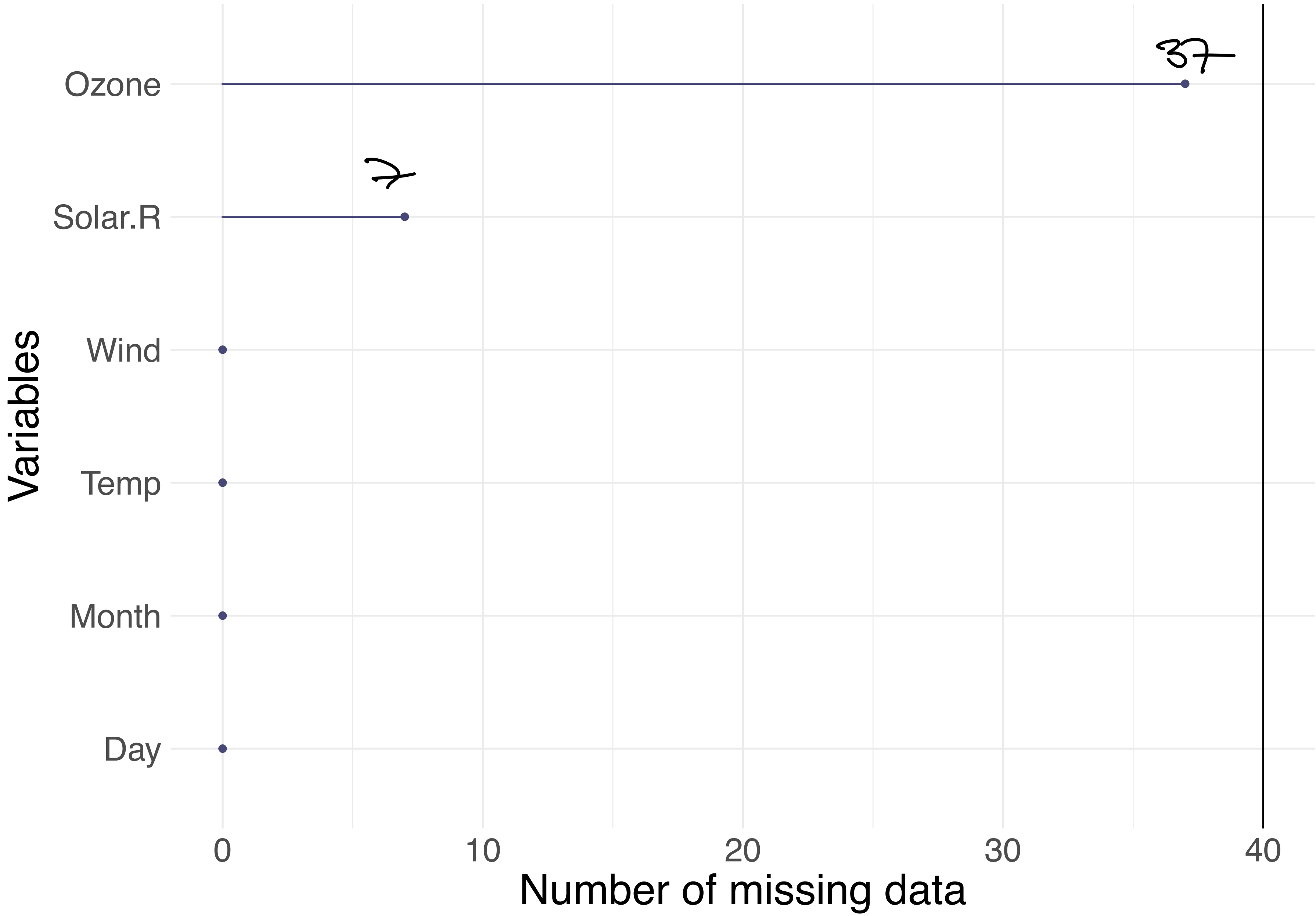
#obs



153

NA's for each variable

```
[1] 153   6
```



37
7
Ozone
Solar.R
Wind
Temp
Month
Day
Variables
0
10
20
30
40
Number of missing data

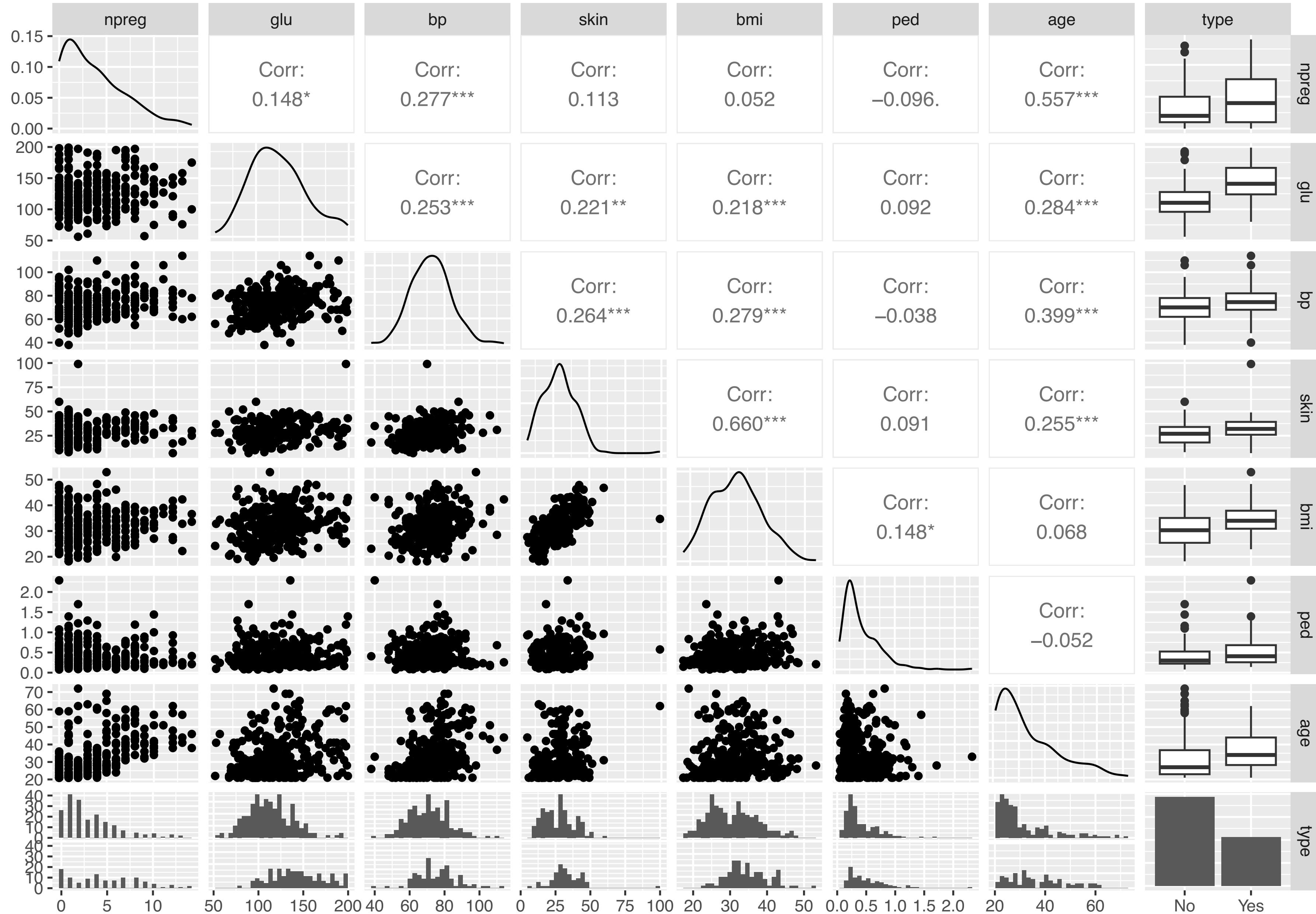
Pima indians

(MASS R package)

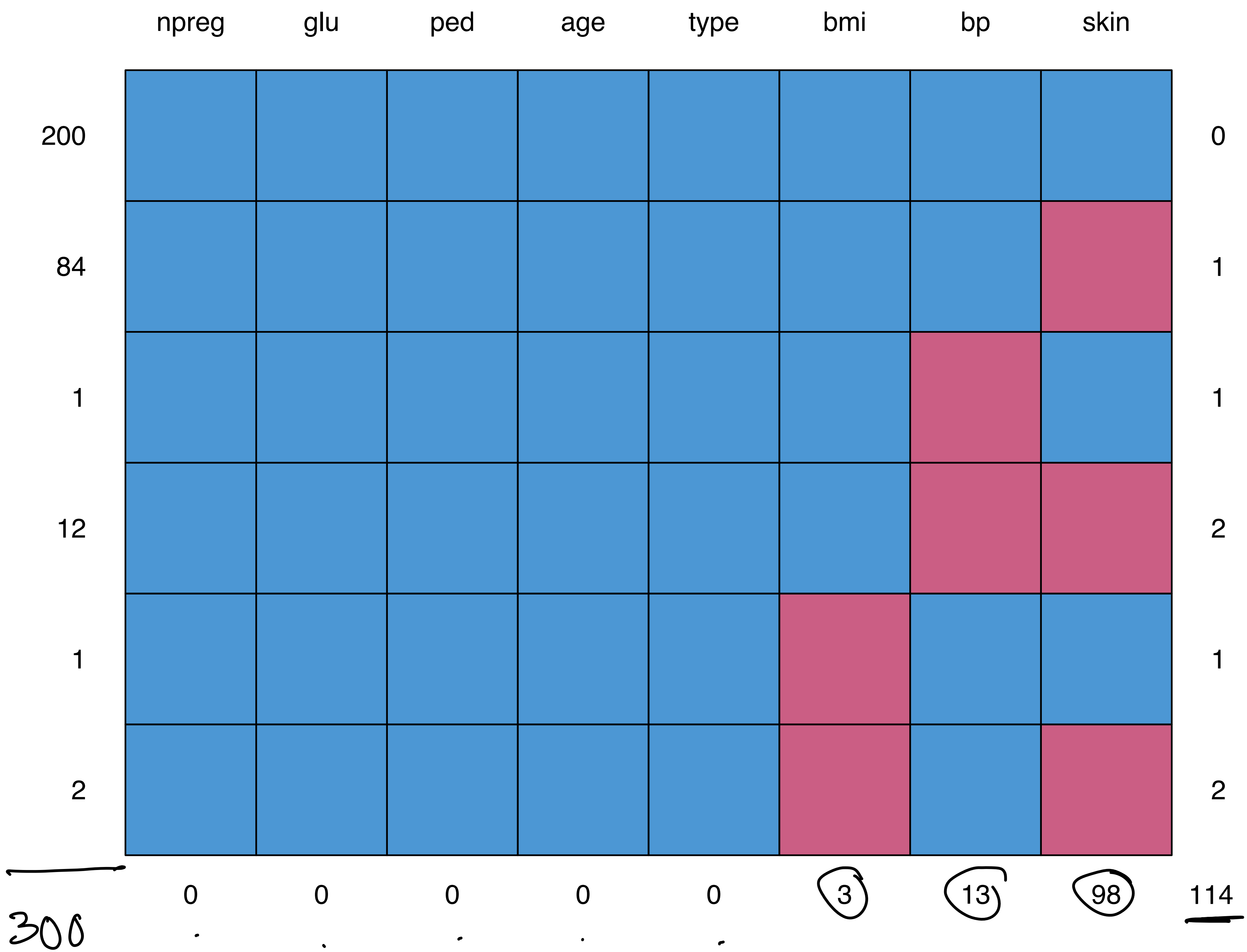
We will use the classical data set of *diabetes* from a population of women of Pima Indian heritage in the US, available in the R `MASS` package. The following information is available for each woman:

- diabetes: 0= not present, 1= present (variable called type)
- npreg: number of pregnancies
- glu: plasma glucose concentration in an oral glucose tolerance test
- bp: diastolic blood pressure (mmHg)
- skin: triceps skin fold thickness (mm)
- bmi: body mass index (weight in kg/(height in m)^2)
- ped: diabetes pedigree function.
- age: age in years

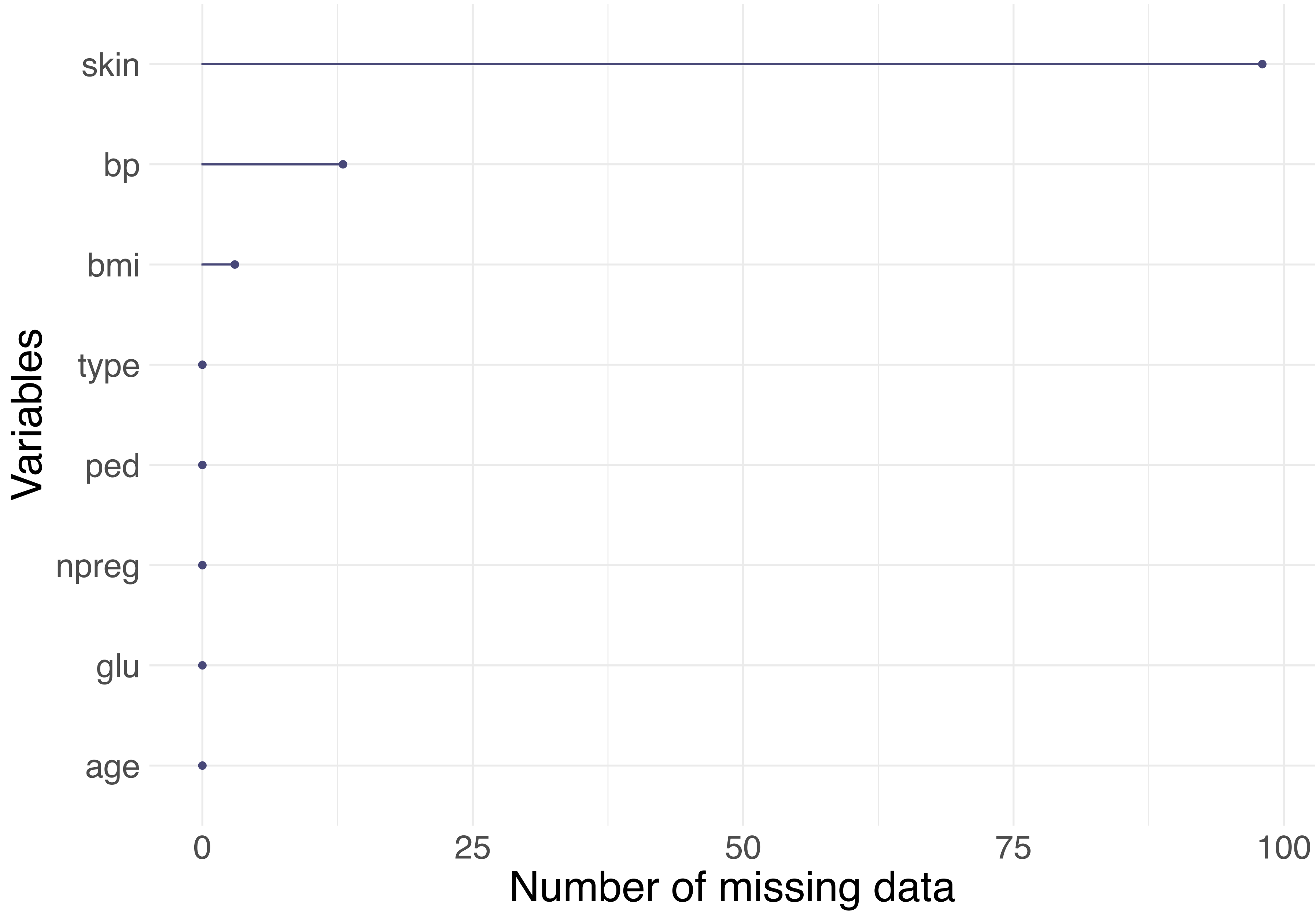
We will look at a data set (Pima.tr2) with a randomly selected set of 200 subjects (Pima.tr), plus 100 subjects with missing values in the explanatory variables.



(Keep in mind: imputation model also needs insight into correlations)


npreg
glu
ped
age
type
bmi
bp
skin
200
84
1
12
1
2
0
1
1
2
1
2
0
0
0
0
0
3
13
98
114
300

```
[1] 300   8
```


skin
bp
bmi
type
ped
npreg
glu
age
Variables
0
25
50
75
100
Number of missing data

Group discussion

Make sure the three types of plots are understood!

- pairs plot
- number of missing values
- missing patterns

Yes - looks good!

Notation

$$
\begin{tikzcd}
 & Z & \\
Y \arrow[ur] & & X \arrow[ul]
\end{tikzcd}
$$

We will use different letters for response and covariates, but often that is not done in other sources. (We will assume that missing values are only present in the covariates and not the response.)

By response we mean the response in the intended *analysis model* and ditto for the covariates. (We will later also talk about an *imputation model* but this is not connected to our notation here.)

Y = response vector

X = covariate matrix $N \times p$

$Z = (X, Y)$

(X_{obs}, X_{mis})

$Z_{obs} = (X_{obs}, Y)$

R = matrix of 0/1

0 = missing

1 = observed

$$\boxed{P(R \mid Z, \psi)}$$

0/1

given observed and missing data + parameter

probability that an obs. is missing

MISSING MECHANISMS

ψ : some parameter in the distribution of R

not related to the parameter of the analysis model

MCAR: missing completely at random

$$P(R \mid Z, \psi) = P(R \mid \psi)$$

all obs have the same prob. of being missing

Not related to X_{obs}, X_{mis}, ψ

$\Rightarrow$ ok to do complete case analysis

Examples:

- measure weight, and the scales run out of battery
- similar mechanism to taking a random sample
- a tube containing a blood sample of study subject is broken by accident and then the blood sample could not be analysed (a set of covariates are then missing)

MAR: missing at random

$$P(R \mid Z, \psi) = P(R \mid Z_{obs}, \psi)$$

probability of missing may depend on observed data (NB also response)

but <u>not dependent on data that is missing</u>!

Examples:

- measure weight, and the scales have different missing proportions when being on a hard or soft surface
- we have a group of healthy and sick individuals (this is the reponse), and for a proportion of the sick individuals the result of a diagnostic test is missing but for the healthy individuals there are no missing values

Most methods for handling missing data require the data to be MAR. If you know that the missingness is at least MAR, then there exists tests to check if the data also is MCAR.

$$\boxed{\text{MNAR = missing not at random}}$$

$$P(R \mid Z, \psi)$$

depend on X_{mis}

"Only" solution: model the missing mechanism

need to be known!

Examples:

- the scales give more often missing values for heavier objects than for lighter objects
- a patient is too sick to perform some procedure that would show a high value of a measurement
- when asking a subject for his/her income missing data are more likely to occur when the income level is high

Group discussion

So far in your study/work/other - you might have analysed a data set (maybe on Kaggle or in a course). Think of one such data set.

- Did this data set have missing values?
- If yes, did you check if the observations were MCAR, MAR or MNAR?
- What did you (or the teacher etc) do to handle the missing data?

If you have not analysed missing data, instead look at the synthetic generation of data with different missing mechanisms below!

SKIPPED IN CLASS —
read for yourself

Synthetic example with missing mechanisms

Example from van Buuren (2018) Chapter 2.2.

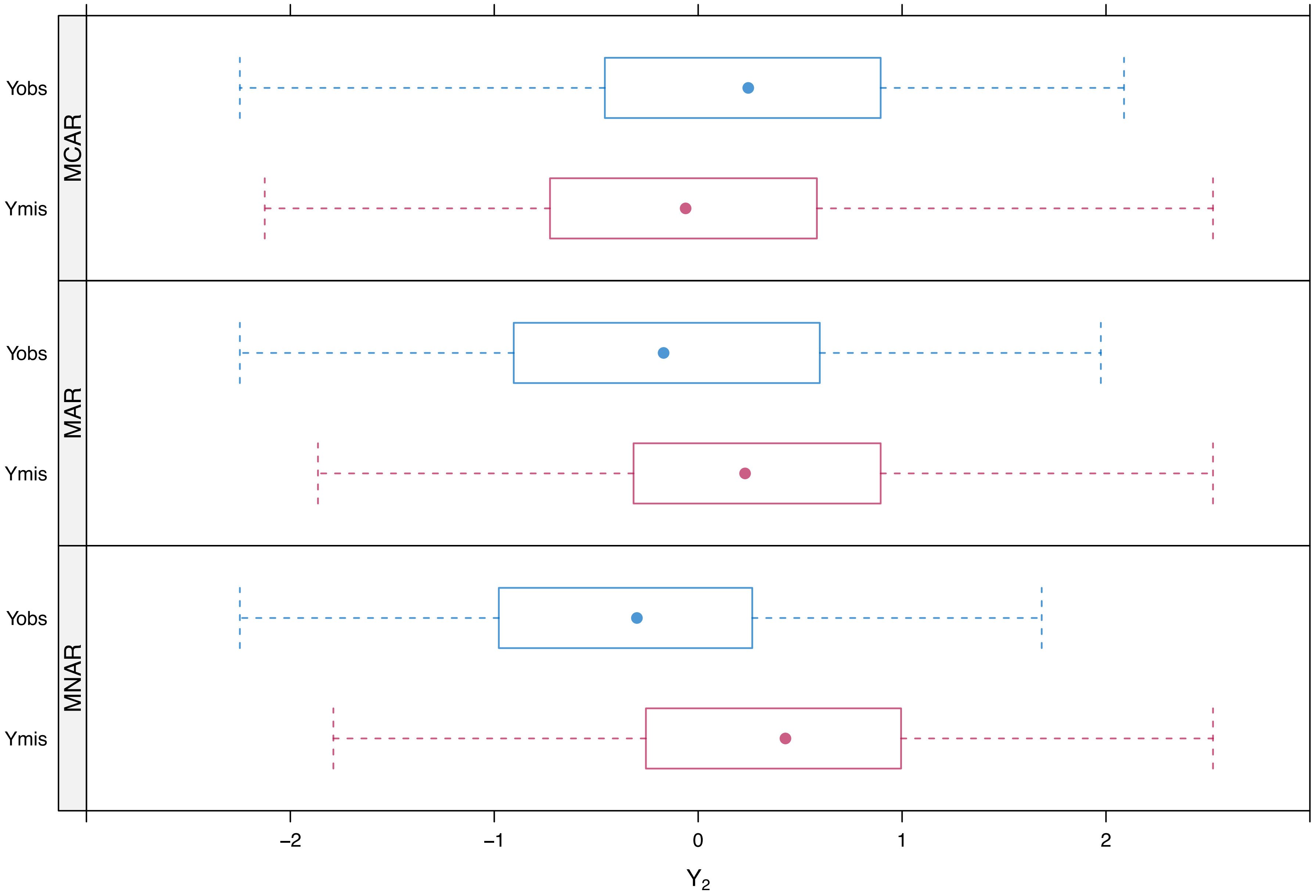
A bivariate (0.5 correlation) normal response (Y_1, Y_2) is generated $N = 300$, and then data are removed from the second component Y_2. This is done in three ways:

- MCAR: each observation Y_2 is missing with probability 0.5
- MAR: each observation Y_2 is missing with probability dependent on Y_1
- MNAR: each observation Y_2 is missing with probability dependent on Y_2.

The boxplots of observed and missing values are shown.

▼ Code

```
#|echo: true
#|warnings: false
#|error: false
# code from https://github.com/stefvanbuuren/fimdbook/blob/master/R/fim
logistic <- function(x) exp(x) / (1 + exp(x))
set.seed(80122)
n <- 300
y <- MASS::mvrnorm(n = n, mu = c(0, 0),
                   Sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
r2.mcar <- 1 - rbinom(n, 1, 0.5)
r2.mar  <- 1 - rbinom(n, 1, logistic(y[, 1]))
r2.mnar <- 1 - rbinom(n, 1, logistic(y[, 2]))
```



MCAR
MAR
MNAR
Yobs
Ymis
Yobs
Ymis
Yobs
Ymis
−2
−1
0
1
2
Y_2

Popular solutions to missing data

Use an analysis method that handles missing data

One such method is the CART classification and regression tree!

How is it done? More in Part 3.

Complete case analysis

Discard all observations containing missing values. This is also called “listwise deletion”.

- Wasteful, but will give valid inference for MCAR.
- If the missing is MAR a complete case analysis may lead to bias. In a regression setting if a missing covariate is dependent on the response, then parameter estimates may become biased.

Let each variable have a probability for missing values of 0.05, then for 20 variables the probability of an observation to be complete is $(1-0.05)^2 0 = 0.36$, for 50 variables 0.08. Not many observations left with complete case analysis. Of cause some variables may have more missing than others, and removing those variables first may of cause lead to less observations that are incomplete

Indicator variable method

Assume we have regression setting with missing values only in one of the covariates. The indicator method generates a new covariate as a missing indicator, and replaces the missing values in the original covariate with 0s.

van Buuren (2018) (Chapter 1.3.7) says that it can be shown that biased estimates of regression parameters can occur also under MCAR. However the method works in particular situation. Which situations this is I (Mette) have not looked into. Would be interesting to know.

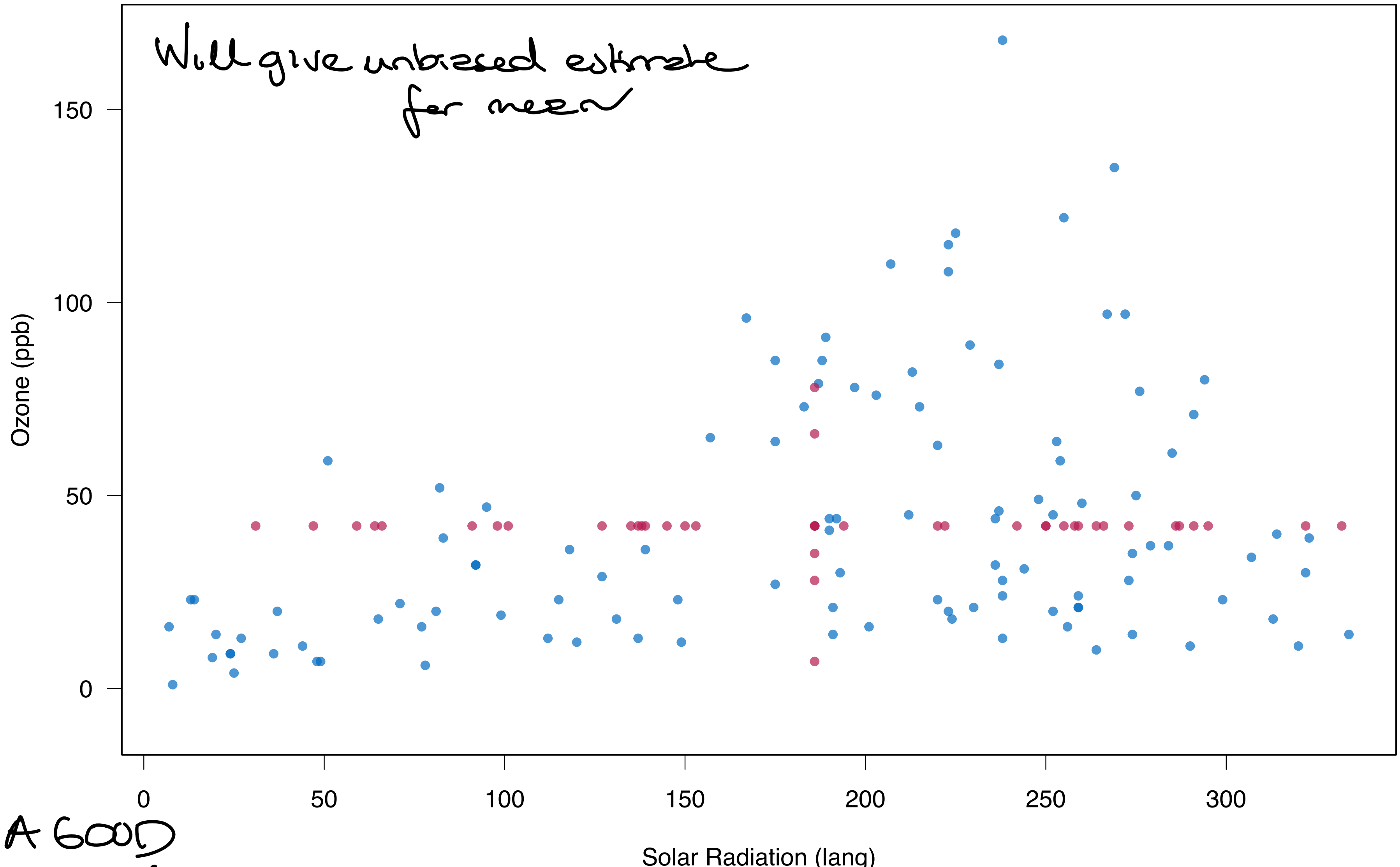
A version of this method is used in machine learning. If the covariate is a categorical covariate then an extra category is created for the missing data. Here more information would be of interest to include!

Maybe something for the article presentation?

Single imputation

here each missing value is imputed (filled in) once by some "estimate" or "prediction" and the data set is then assumed to be complete and standard statistical methods are used.

MEAN imputation : fill in average over all observations

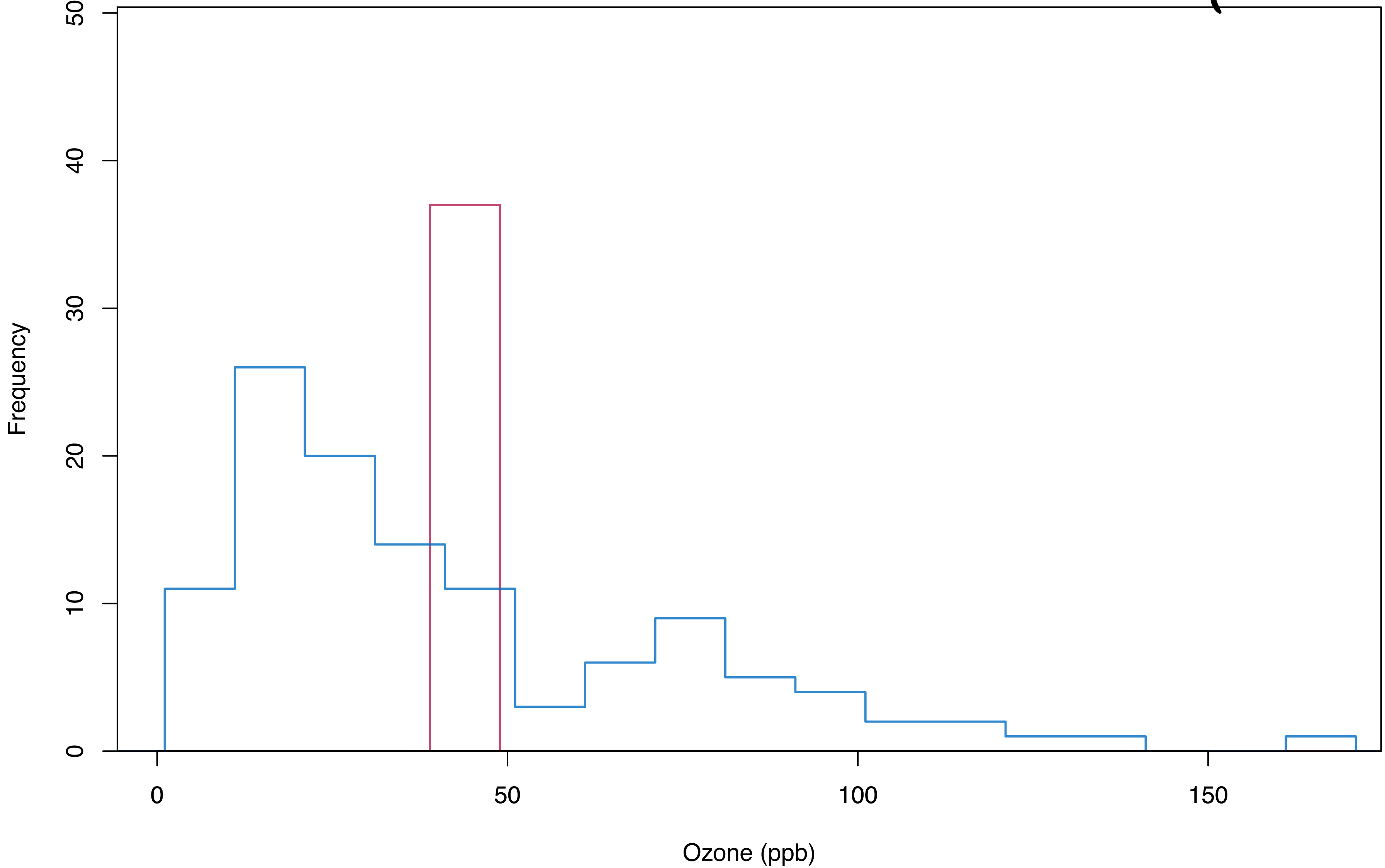


NOT A GOOD SOLUTION

blue : observed
red : imputed

How does this new imputed differ from original?
- less variable
- trends may be blurred

distribution of
obs & imp



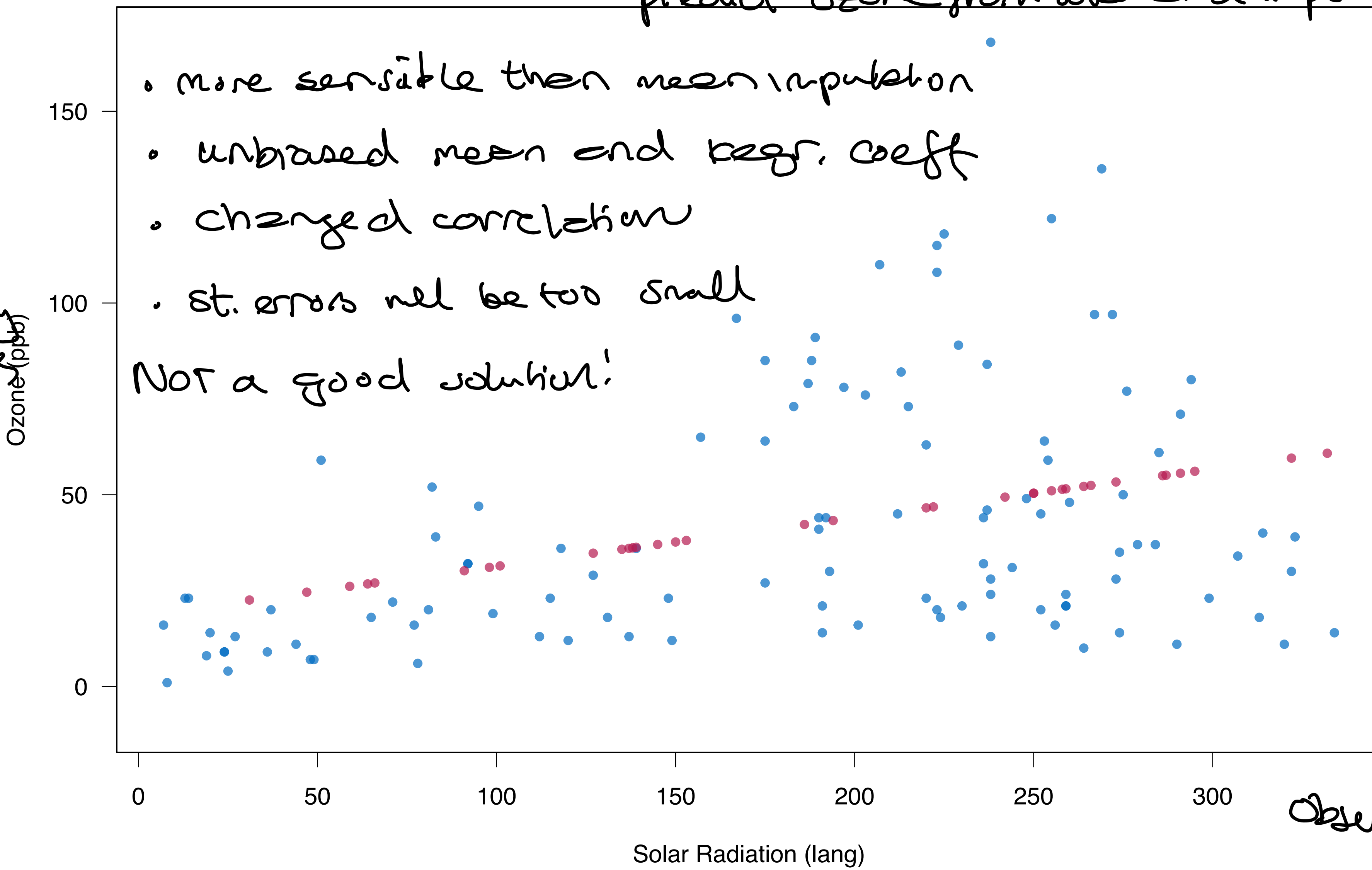
Quality of mean imputation: mean unbiased under MCAR, and regression weights or correlations not. Standard errors too small.

REGRESSION imputation

- estimate regression line on complete observations
- predict ozone from solar and impute

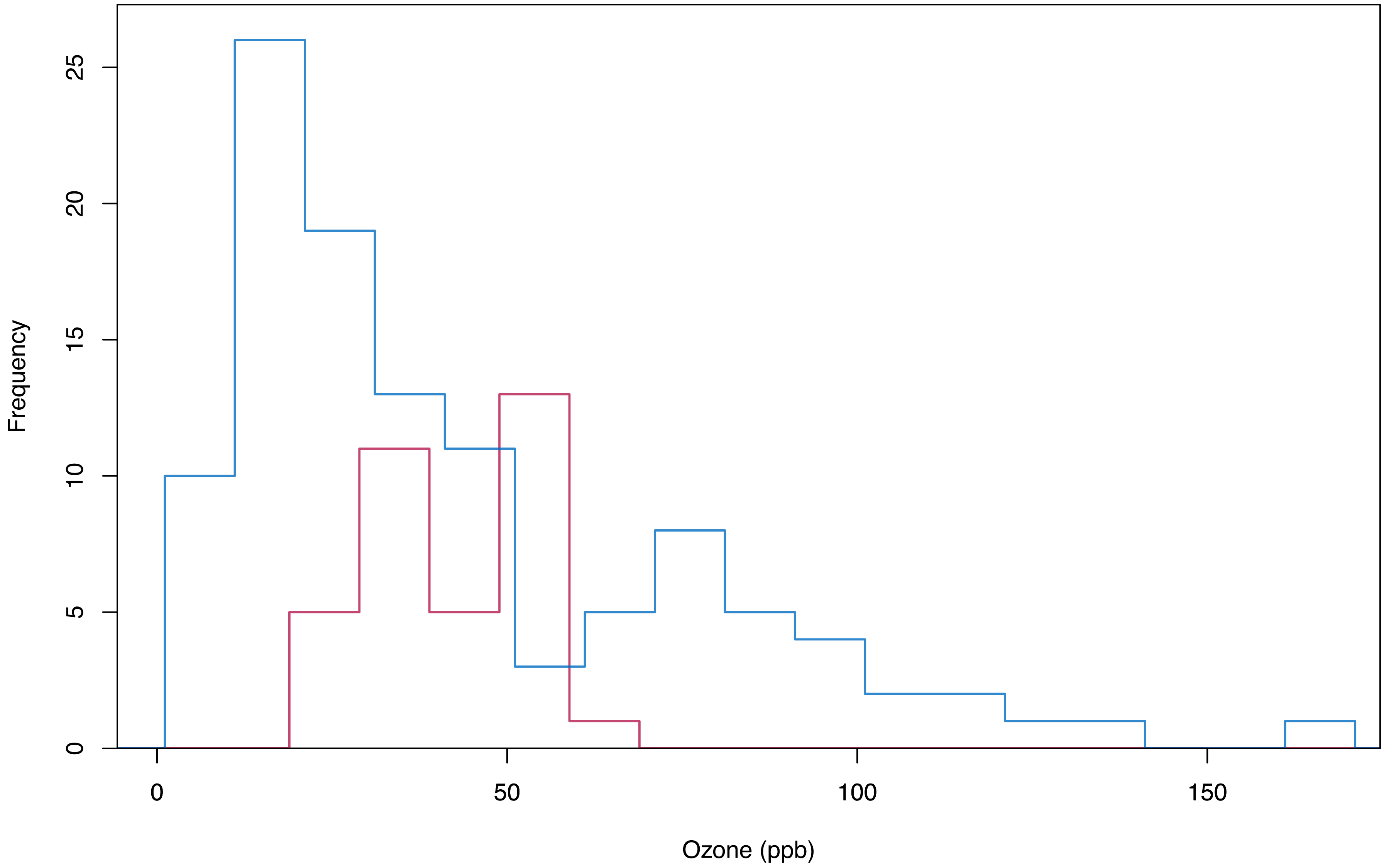
- more sensible than mean imputation
- unbiased mean and regr. coeff
- changed correlation
- st. errors will be too small

NOT a good solution!



partly missing

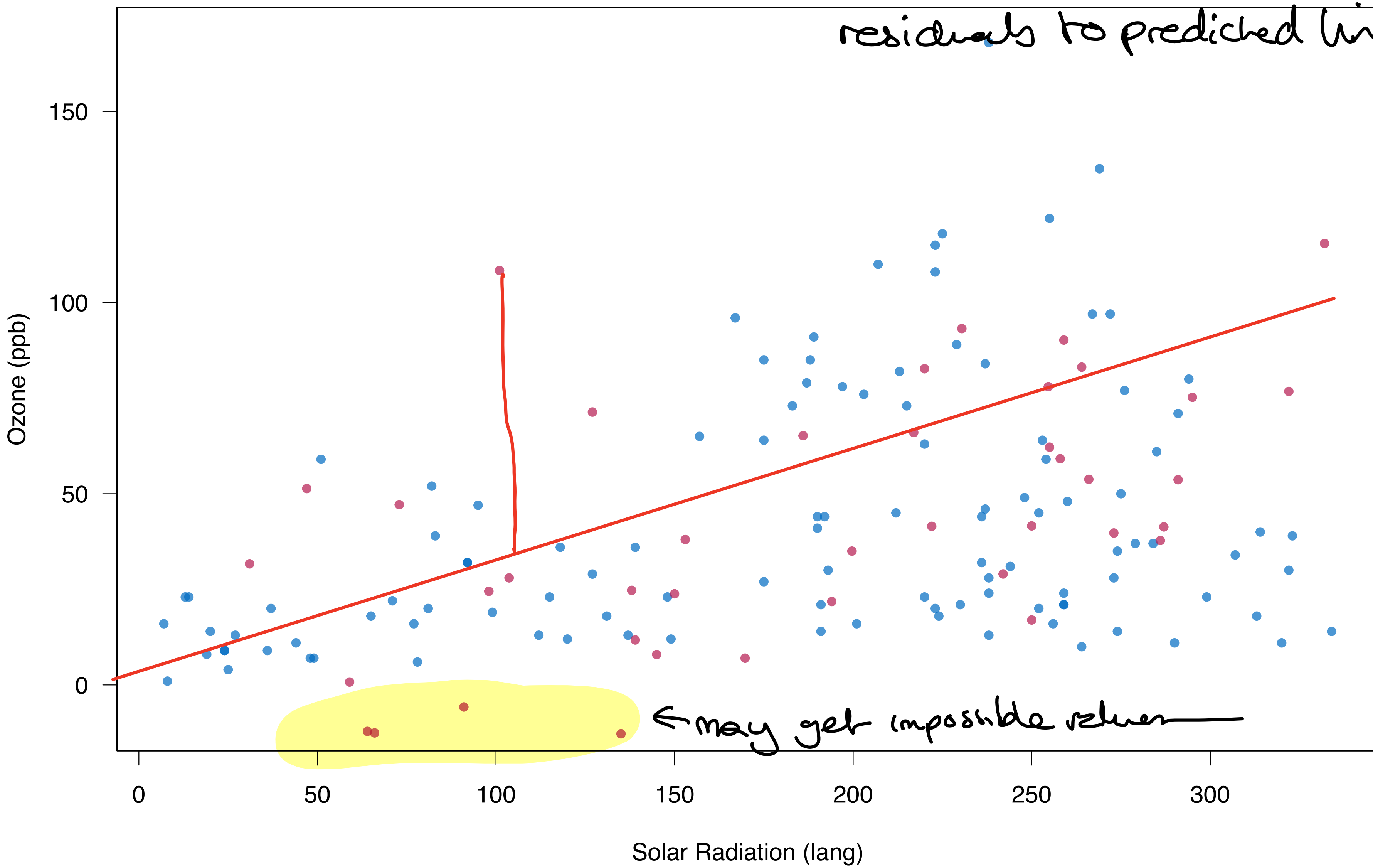
Observed


Frequency
0
5
10
15
20
25
0
50
100
150
Ozone (ppb)

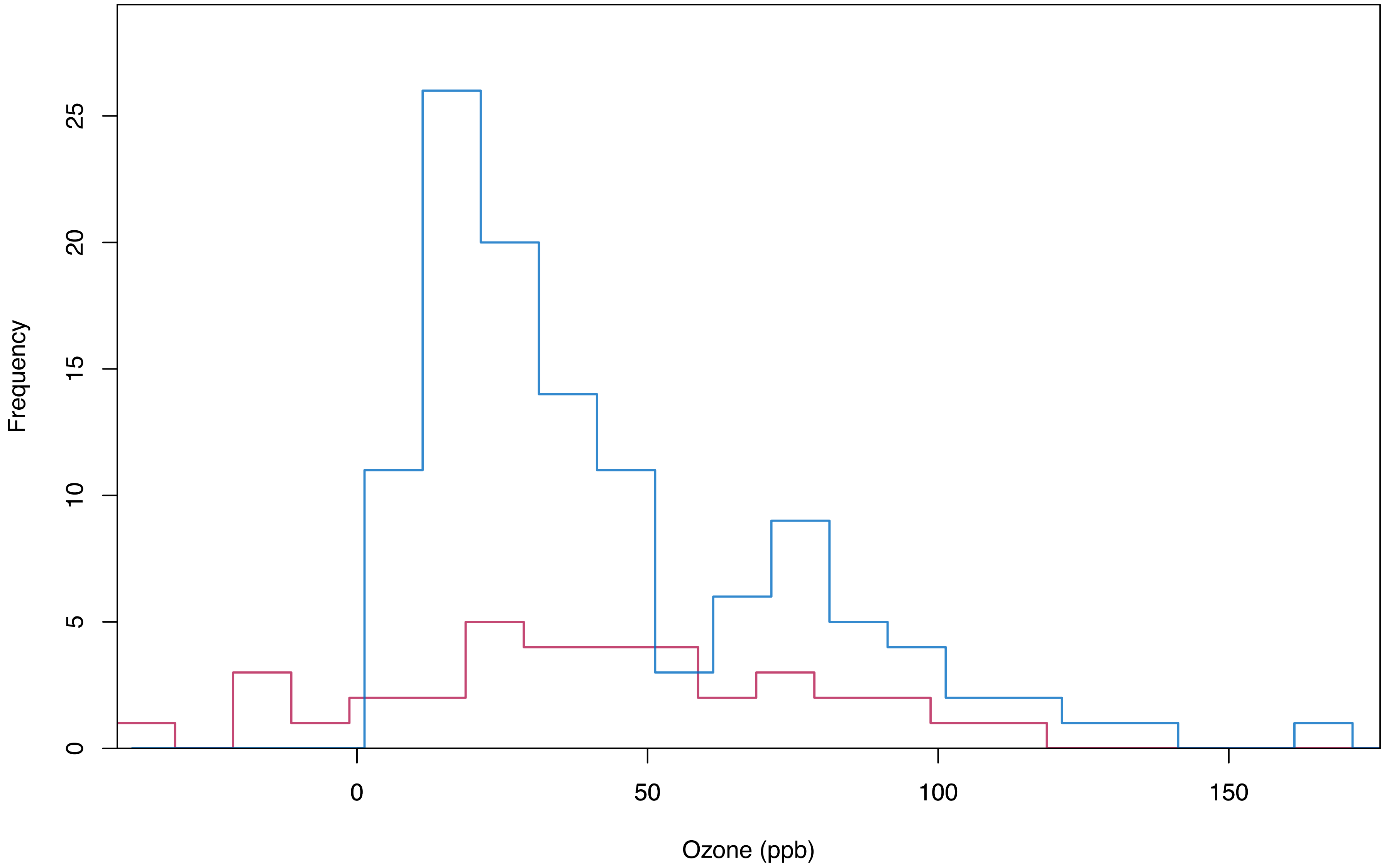
Quality of regression imputation: mean and regression weights are unbiased under MAR. Correlation is not. The imputed red dots have correlation 1 (linear relationship). Standard errors too small.

STOCHASTIC regression imputation

- fit regression complete obs
- add random draw from residuals to predicted line



← may get impossible values



Frequency
0
5
10
15
20
25
0
50
100
150
Ozone (ppb)

Quality of stochastic regression imputation: mean, regression weights and correlations are unbiased under MAR. Standard errors too small.

Group discussion

Of the single imputation methods the stochastic regression imputation method appears to be the best. Do you see why? Would you think of possible improvements to this method?

- nonlinear regressions
- multiple regression – not only simple
- draw from distribution of $\hat{\beta}$ at x, not residuals
- not one imputed dataset – but many!

Likelihood approaches

(Not included in 2023)

For example

- Bjørnland et al. Extreme phenotype sampling
- EM-algorithm from TMA4300

Fully Bayesian approaches

Sadly, not covered here.

Received: 9 January 2017 | Revised: 20 June 2018 | Accepted: 25 June 2018
DOI: 10.1002/sim.7914

Model the missingness!

RESEARCH ARTICLE

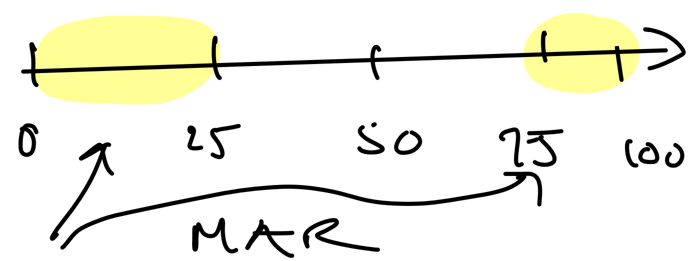
WILEY Statistics in Medicine

Powerful extreme phenotype sampling designs and score tests for genetic association studies

FTO

WHR
BMI

Thea Bjørnland[1] | Anja Bye[2] | Einar Ryeng[3] | Ulrik Wisløff[2] | Mette Langaas[1]



sampling strategy

2.3 | Extreme phenotype sampling

We define a general extreme phenotype sampling design where the classification rule (extreme or not extreme) can be tailored to each individual in the sample.

Definition 5. (**Extreme phenotype sampling**)
Individual i has an extreme phenotype if $Y_i < l_i$ or $Y_i > u_i$, where l_i and u_i are known thresholds. All individuals who are classified as extreme are selected for genotyping.

3.1 | Complete case analysis

Using the conditional phenotype distribution $Y_i|(Y_i < l_i \cup Y_i > u_i)$, where classification rules l_i and u_i are determined before seeing the data, the likelihood for the complete cases (Definition 3) is

$$L_C = \prod_{i \in C} \frac{\frac{1}{\sigma}\phi\left(\frac{Y_i-\mu_i}{\sigma}\right)}{1-\Phi\left(\frac{u_i-\mu_i}{\sigma}\right)+\Phi\left(\frac{l_i-\mu_i}{\sigma}\right)},$$

Multiple imputation

Short historical overview

Historically multiple imputation dates back to Donald B. Rubin in the 1970´s. The idea is that multiple data set (multiple imputations) will reflect the uncertainty in the missing data. To construct the m data sets theory from Bayesian statistics is used, but executed within the frequentist framework. Originally $m = 5$ imputed data sets was the rule of thumb.
The method did not become a standard tool until 2005 (according to van Buuren (2018), 2.1.2), but now in 2023 it is widely used in statistics and has replaced version of single imputation. However, multiple imputation is not main stream in machine learning.

STEPS of multiple imputation

$Z = \{ X_{obs}, X_{mis}, Y \}$ our data

In the analysis model we aim to relate Y to X

slim
bmi

ozone: $Y = X\beta + \varepsilon$ diabetes: $Y_i \sim bin(1, p_i)$ $\text{logit } p_i = x_i^T \beta$

wind   ozone   0/1

1) Devise an imputation model (often of regression type) where each missing covariate is modelled as a function of other covariates and the analysis method response and possibly other variables. → Make m full data sets

Of interest is Q (or just β) in the analysis model, but in the imp. model we have other parameters (γ's?)

2) Analyze the m complete data sets (know how to do that)

$\hat{\beta}$, $\hat{Cov}(\hat{\beta})$

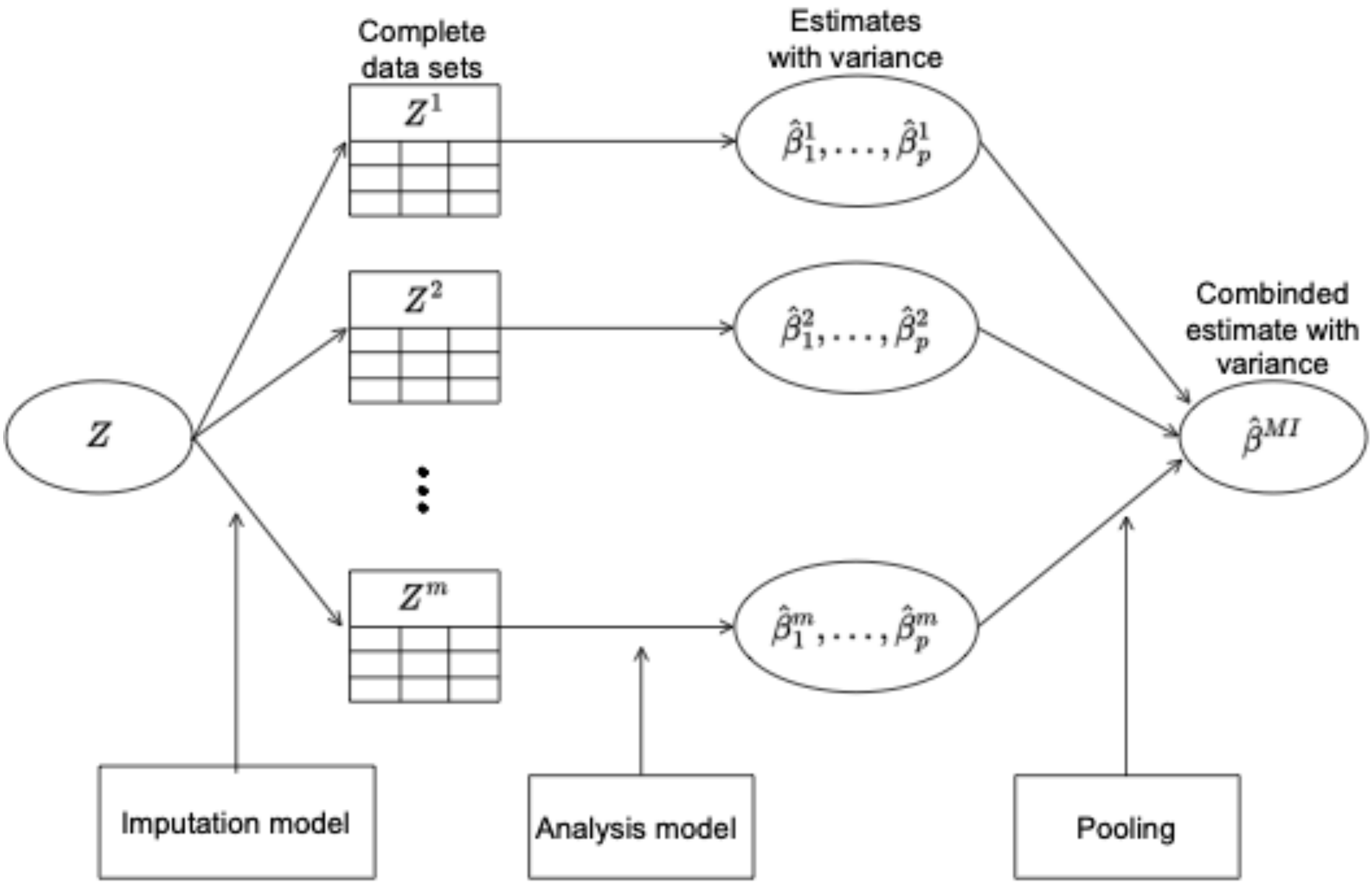
$\hat{Q}$ $Cov(\hat{Q})$ $\Rightarrow$ m results

3) Combine the results using a set of rules: <u>Rubin's rules</u>

4) Use the results directly or indirectly

IMPORTANT: the m data set are not to be used as complete datasets per se $\rightarrow$ they are only used to estimate Q or β with uncertainty!

Schematic for multiple imputation from Marthe Bøe Ludvigsen project thesis.



Rubin's rules

Algorithmic view

Fist we look at formulas for our quantities of interest, and next the Bayesian motivation for the formulas.

Quantity of interest

We denote our quantity (parameter) of interest by $\mathbf{Q}$, and assume this to be a $k \times 1$ column vector.

Example 1: Multiple linear regression

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Here $\mathbf{Q} = \beta$.

Example 2: Logistic regression

Again $\mathbf{Q} = \beta$.

$\mathbf{Q}$ can also be a vector of population means or population variances. It may not depend on a particular sample, so it cannot be a sample mean or a p-value.

Estimators: for each of the m data sets $l = 1, .., m$

$\hat{Q}_l$ is this estimator: Ex 1: $\hat{\beta}_l = (X_l^T X_l)^{-1} X_l^T y_l$ windusozone+

Ex 2: $\hat{\beta}_l$ again - not closed form

diabetes

POOLED ESTIMATOR:

$$\overline{Q} = \frac{1}{m} \sum_{l=1}^{m} \hat{Q}_l$$

Rubins rule for Q

Intuitive and simple!

VARIANCE OF ESTIMATOR

Let $\bar{U}_l$ be the $\widehat{Cov}(\hat{Q}_l)$

Ex1: $(X_l^T X_l)^{-1} \hat{\sigma}_l^2$

Ex2: Inverse Fisher info

1) Within imputation variance

$$\bar{U} = \frac{1}{m} \sum_{l=1}^{m} \bar{U}_l$$

2) Between imputation variance

$$B = \frac{1}{m-1} \sum_{l=1}^{m} (\hat{Q}_l - \bar{Q})(\hat{Q}_l - \bar{Q})^T$$

3) Total variance of $\bar{Q}$: T

$$T = \bar{U} + B + \frac{B}{m} = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

3) Total variance of $\overline{\mathbf{Q}}$

$$\mathbf{T} = \overline{\mathbf{U}} + \mathbf{B} + \frac{\mathbf{B}}{m} = \overline{\mathbf{U}} + (1 + \frac{1}{m})\mathbf{B}$$

- First term: variance due to taking a sample and not examining the entire population (our conventional variance of estimator.
- Second term: extra variance due to missing values in the samples
- The last term is the simulation error: added because $\overline{\mathbf{Q}}$ is based on finite m

Friday: Bayesian interpretation $E(E(X|Y))$ + tot var

→ W1 Ex1 ← look at

HOMEWORK:

+ Handbook 12.2