COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Missing Data

- Missing values: aka most common type of data
- In R they are denoted NA
- Why do we care?

Simple example in R:

```
x = c(4, 6, 2)
mean(x)
[1] 4
x = c(4, 6, NA)
mean(x)
[1] NA
```

Overview of Methods

Methods that throw away data (older approaches, prior to 1987

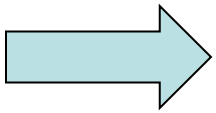
- Complete cases analysis aka CCA (not a great idea)
- Complete variables (bad idea)
- Pairwise deletion (for special purposes)

Methods that don't throw away data (new approaches, 1990s to present)

- Imputation: single and multiple

CCA

X_1	X_2	X_3	X_4
0	2	5	2
0	1	3	NA
0	2	4	3
0	3	5	NA
1	5	NA	NA
1	4	NA	12
1	6	11	NA
1	6	12	16



X_1	X_2	X_3	X_4
0	2	5	2
0	2	4	3
1	6	12	16

Complete cases (aka listwise deletion)

- Removes all observations from the dataset that have any missing values (most software packages do this automatically when you run an analysis, for example, the `lm` function in R)
- At best it is inefficient (yields higher standard errors) because of reduced sample size
- At worst it can cause severe bias
- Reductions in sample size may preclude certain types of analyses, e.g. subgroup analyses
- When does it work?

Missing Data Mechanisms

Rubin (1976) classified missing data into 3 categories:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR), also called Missing Not at Random (MNAR)
- Aka the most confusing statistical terms ever invented

The process that governs the probability of a data point being missing is called *missing data mechanism* or response mechanism.

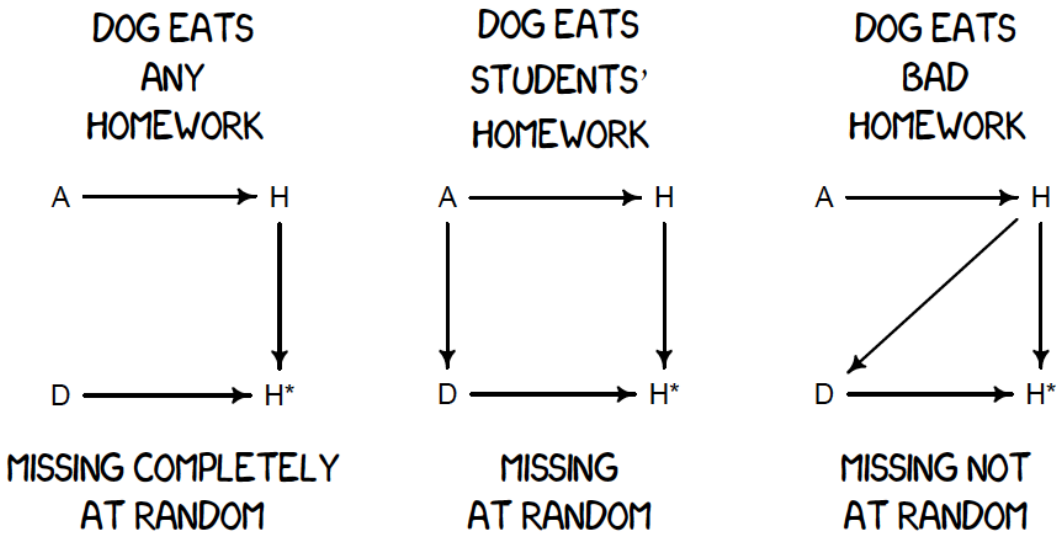
MDM: Example

Consider a cancer study in which one of the variables is a quality-of-life score, which has missing entries.

- If missing entries are due to a random computer glitch or human error, unrelated to the patients at all, then we have MCAR
- If missing quality of life is a function of age and education, or treatment and health status, which we observe directly, then we have MAR
- If missing is direct function of quality of life, then we have NMAR

- H: Homework
- H*: Homework with missing values
- A: Attribute of student (say, study time)
- D: Dog (missingness mechanism)

Three Types of Missingness



Notation

- Sample data matrix is usually denoted $\boldsymbol{X}$, that is, the $n \times p$ matrix containing the data values on p variables for all n units in the sample.

- We define the *response indicator* $\boldsymbol{R}$ as an $n \times p$ 0–1 matrix (see next slide for details).

- Specific elements of $\boldsymbol{X}$ and $\boldsymbol{R}$ are denoted by y_{ij} and r_{ij}, respectively.

- We are restricted to the case where $\boldsymbol{R}$ is completely known, i.e., we know where the missing data are. This covers many applications of practical interest, but not all. For example, some questionnaires present a list of diseases and ask the respondent to place a "tick" at each disease that applies. If there is a "yes" we know that the field is not missing. However, if the field is not ticked, it could be because the person didn't have the disease (a genuine "no") or because the respondent skipped the question (a missing value).

Notation

Let R be the **matrix** of variables $R_1, \ldots, R_p$, corresponding to the variables in our dataset, $X_1, \ldots, X_p$, that indicate whether a given value of the corresponding X variable is observed (= 1) or missing (= 0)

X_1	X_2	X_3	X_4
0	2	5	2
0	1	3	?
0	2	4	3
0	3	5	?
1	5	?	?
1	4	?	12
1	6	11	?
1	6	12	16

R_1	R_2	R_3	R_4
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	0
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

Missing Completely at Random

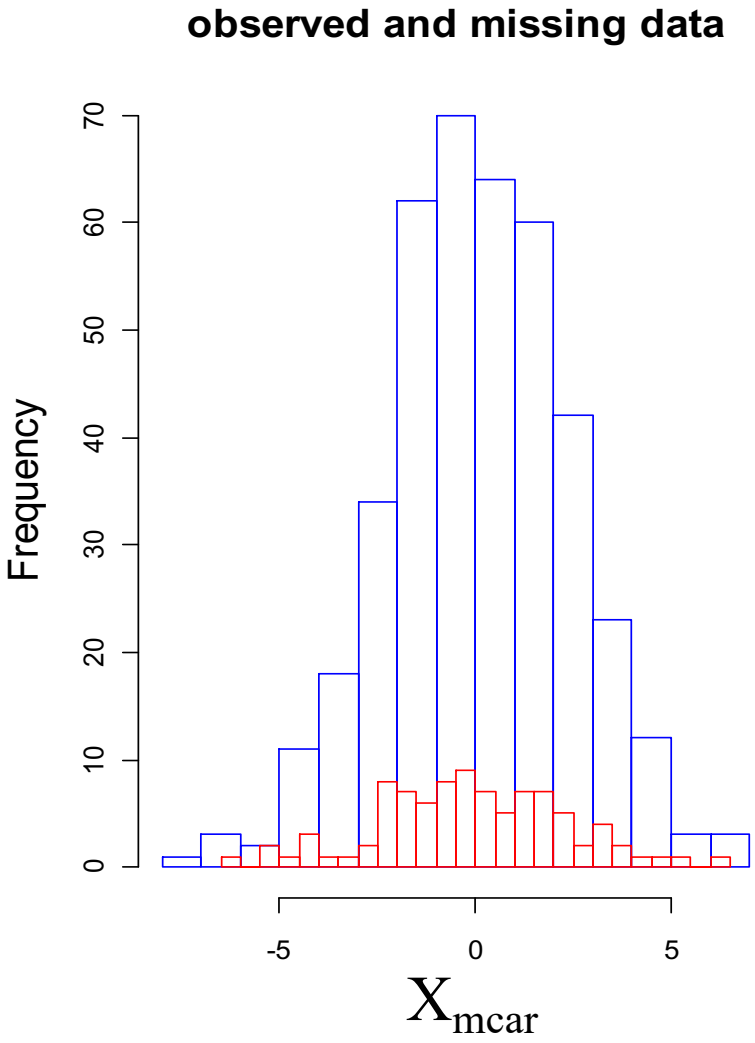
- More formally, let $\boldsymbol{R}$ be the missing pattern and $\boldsymbol{X}$ the data
- In MCAR $\boldsymbol{R}$ and $\boldsymbol{X}$ are assumed independent random vectors ($\boldsymbol{R} \perp \boldsymbol{X}$)
- $P(R_1, R_2, \ldots, R_p \mid X_1, X_2, \ldots, X_p) = P(R_1, R_2, \ldots, R_p)$
- $P(\boldsymbol{R} = \mathbf{0} \mid \boldsymbol{X}_{obs}, \boldsymbol{X}_{mis}, \varphi) = P(\boldsymbol{R} = \mathbf{0} \mid \varphi)$

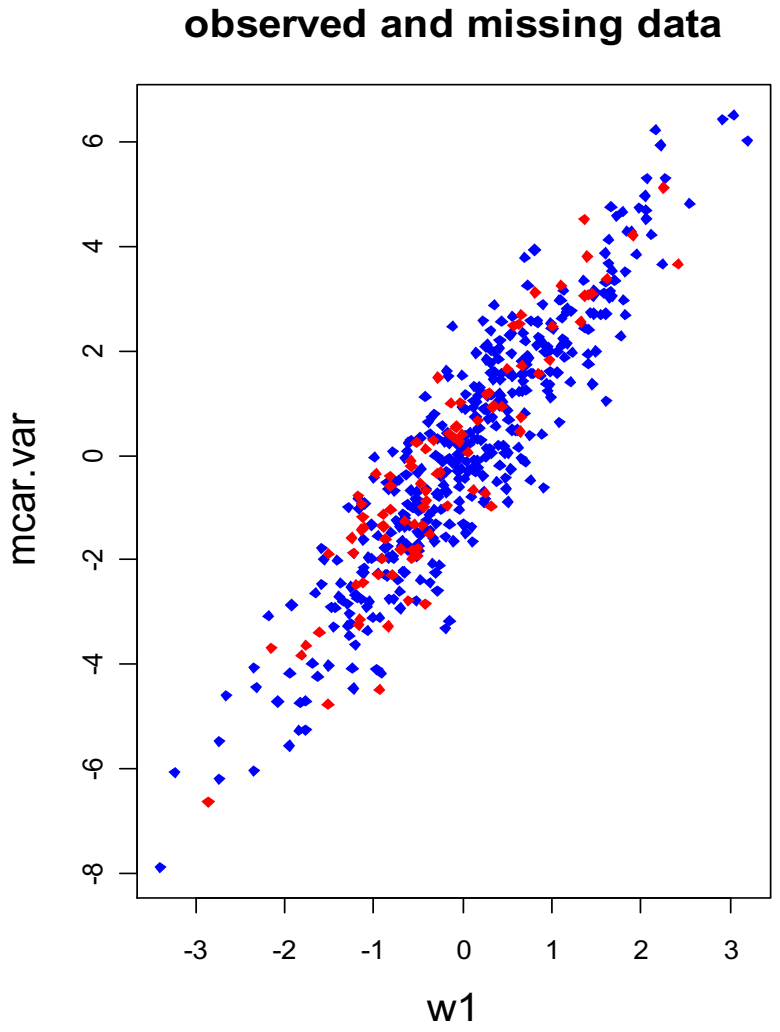
- This means that whether any given value is missing is *completely random* (doesn't depend on any other variable).
- This is generally not a plausible assumption. Usually, certain types of people/patients are much more likely than others to have missing data.
- When could this happen in reality?
  - A bunch of records are lost
  - Missing by design

MCAR data (if we could observe it!):

blue = observed

red = missing





Estimation with MCAR

- Suppose we are interested in estimating the mean of the population $\mu = \mathrm{E}(X)$, but instead of iid sample $X_1, \ldots, X_n$, we observe X_i if $R_i = 1$, and X_i is missing when $R_i = 0$.
- The complete cases estimator is:

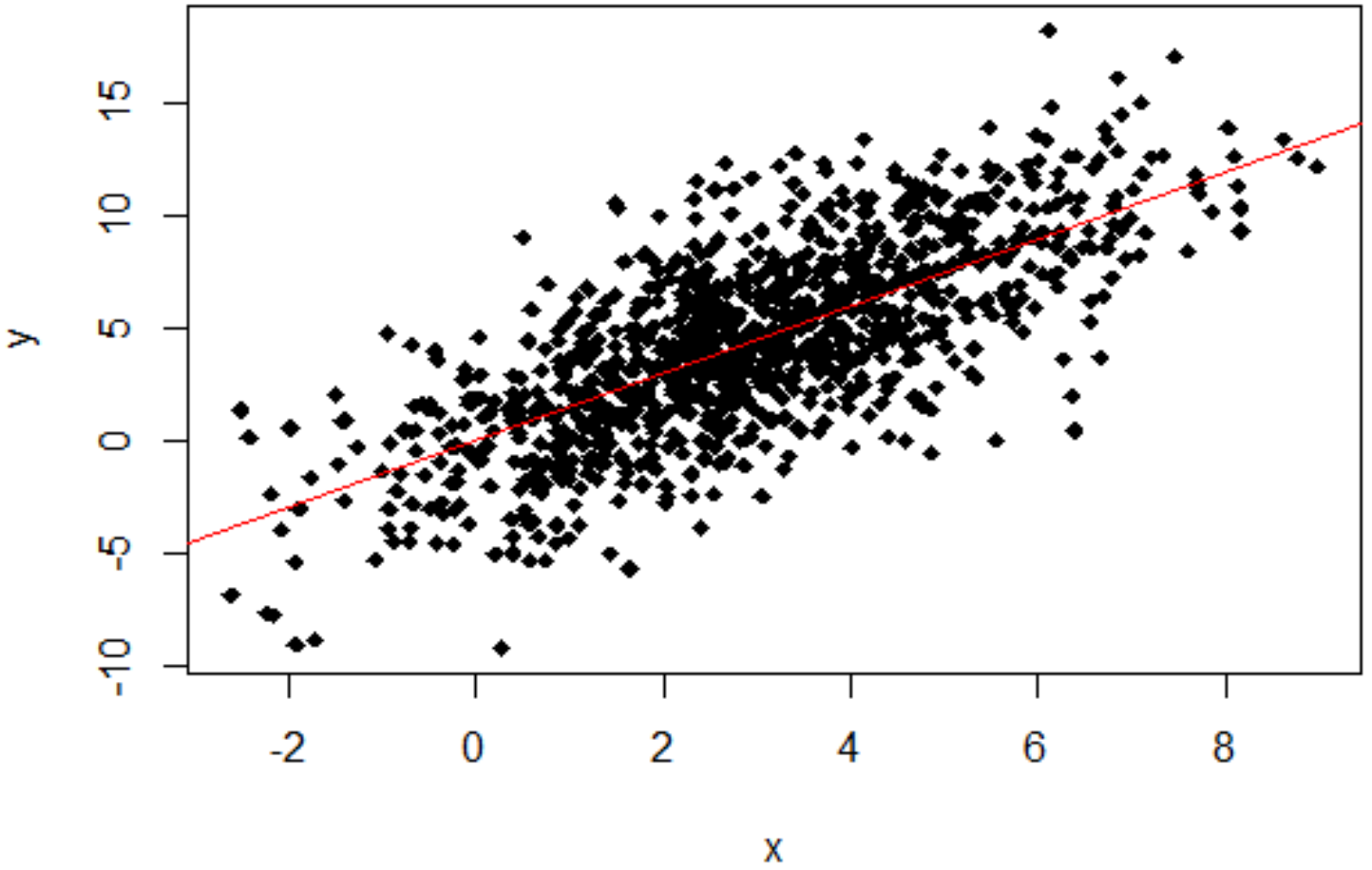
$$\hat{\mu}^C = \frac{\sum_{i=1}^n R_i X_i}{\sum_{i=1}^n R_i}$$

- Assuming MCAR, that is, $\mathrm{P}(R = 1 \mid X) = \mathrm{P}(R = 1)$, it can be shown that

$$\hat{\mu}^C = \frac{\frac{1}{n}\sum_{i=1}^n R_i X_i}{\frac{1}{n}\sum_{i=1}^n R_i} \xrightarrow{P} \frac{E(RX)}{E(R)} = \frac{E(R)E(X)}{E(R)} = E(Y) = \mu$$

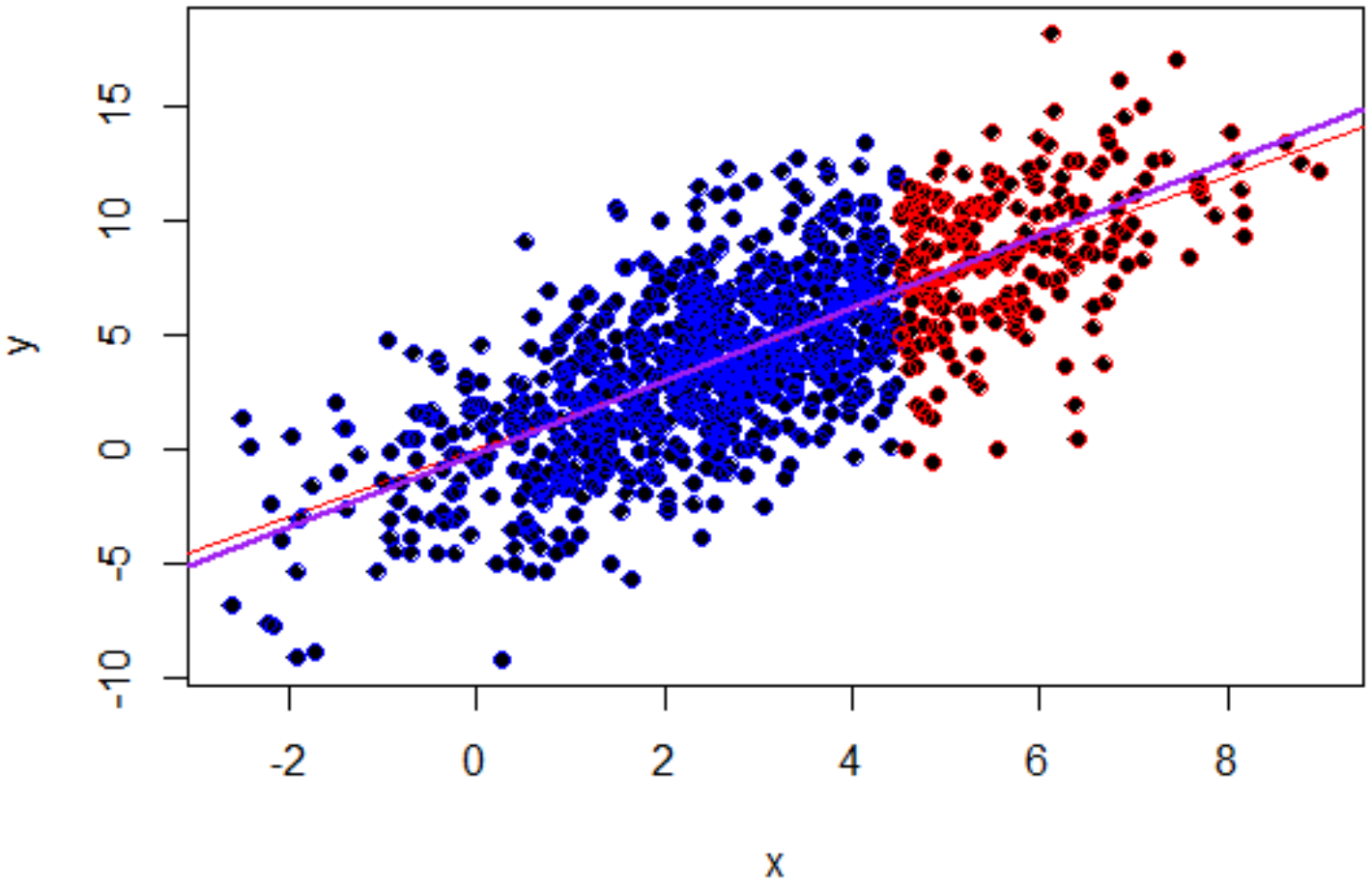
Illustration of CCA and

.



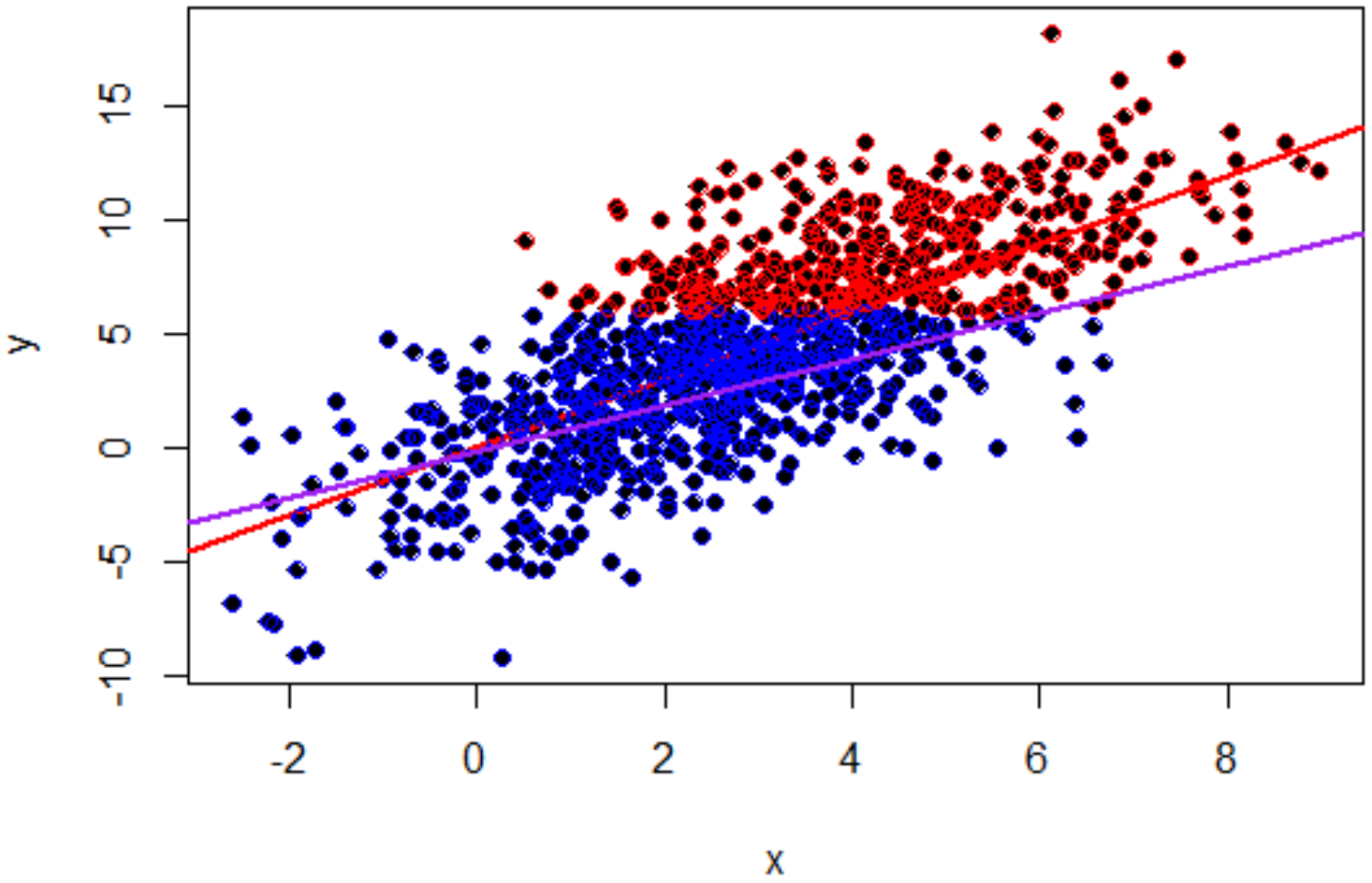
CCA and regression: just fitting to observed (blue) points

Missingness due to x



CCA and regression: just fitting to observed (blue) points

Missingness due to y



Pairwise Deletion

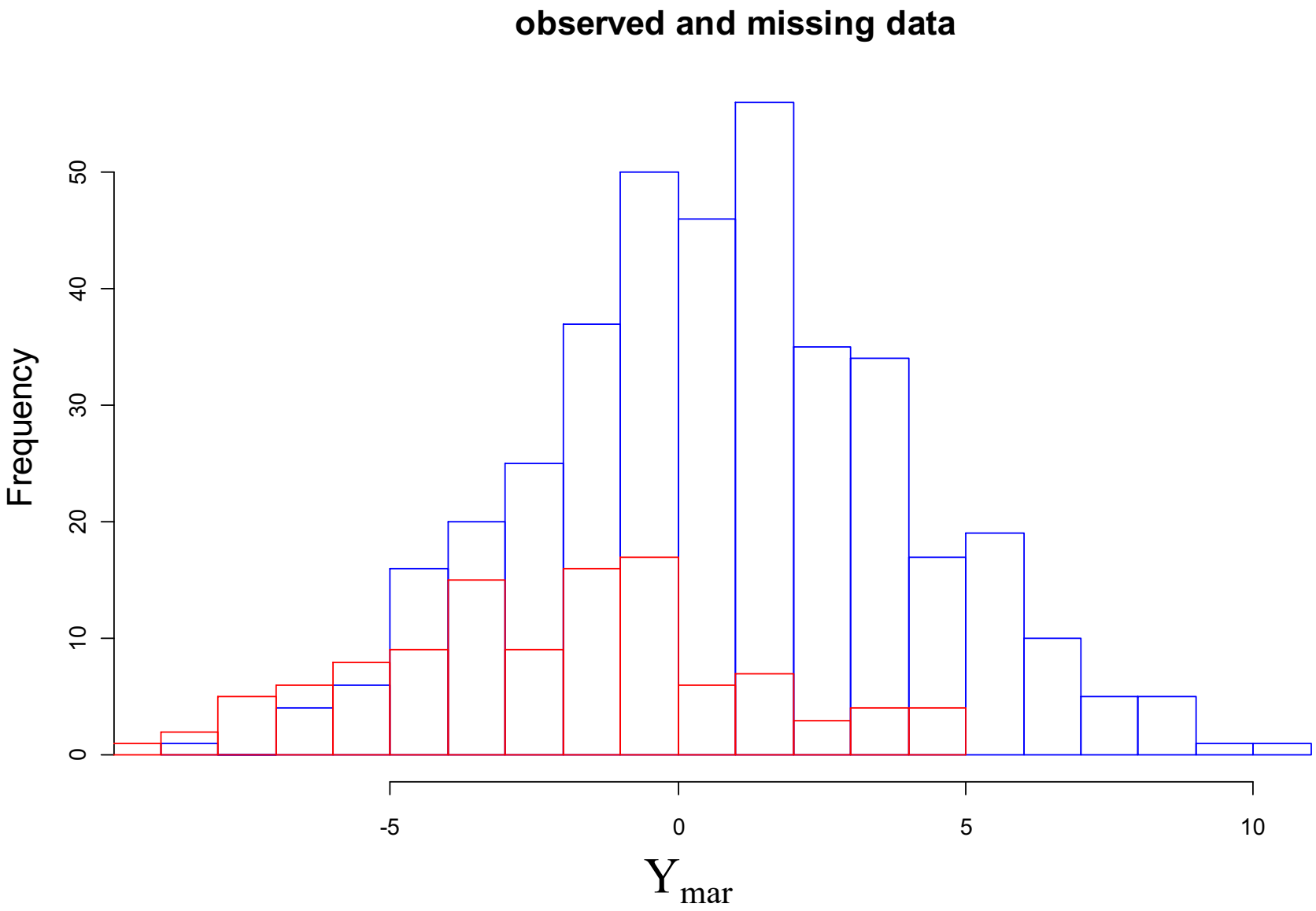
- Procedure that focuses on the covariance matrix.
- Each element of that matrix is estimated from all data available for that element.
- Because different variances and covariances are based on different subsamples of respondents, parameter estimates may be biased unless missingness is MCAR.
- (More technical): because the different parameters are estimated with different subsamples, it often happens that the matrix is not positive definite!

Missing at Random (MAR)

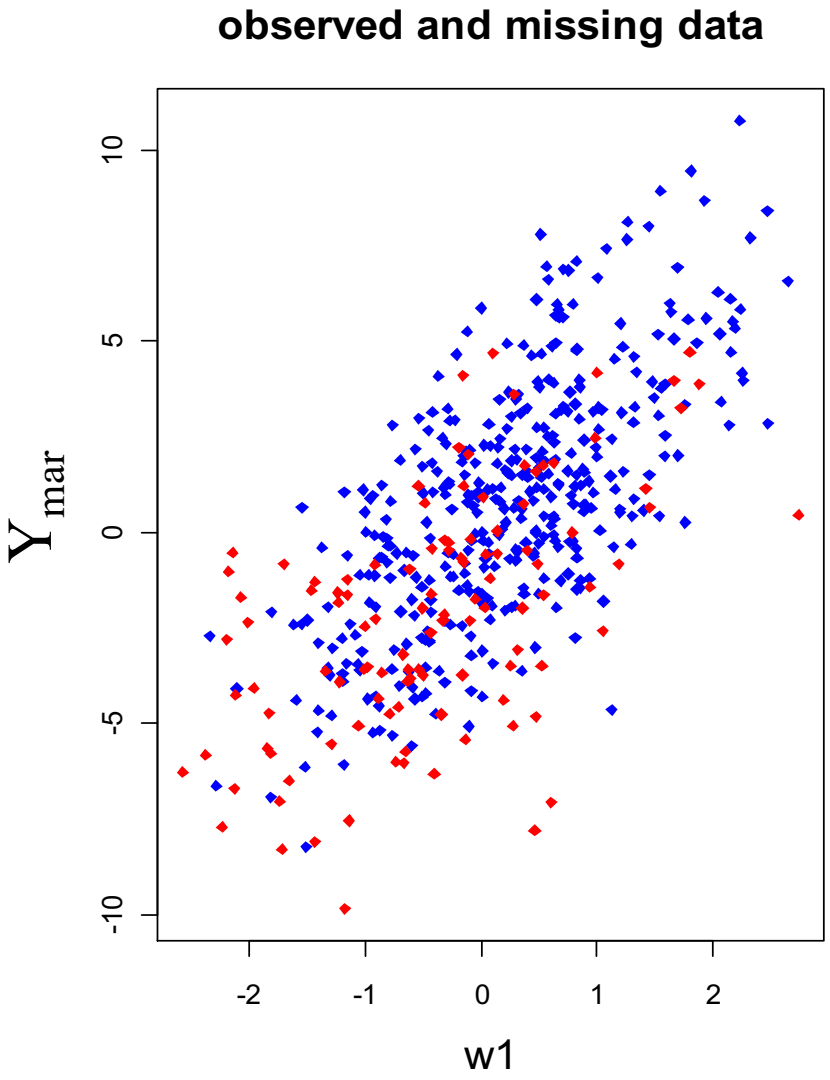
- The propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
- $P(R_1, \ldots, R_p \mid X_1, \ldots, X_p) = P(R_1, \ldots, R_p \mid X_1^{obs}, \ldots, X_p^{obs})$
- $P(\boldsymbol{R} \mid \boldsymbol{X}) = P(\boldsymbol{R} \mid \boldsymbol{X}^{obs})$
- Here missingness depends on observed values of the variables.
- A simple version of this is $P(R_1 \mid X_1, \boldsymbol{W}) = P(R_1 \mid W)$, where $\boldsymbol{W}$ is a subset of fully observed variables in the matrix $\boldsymbol{X}$
- Classic example:

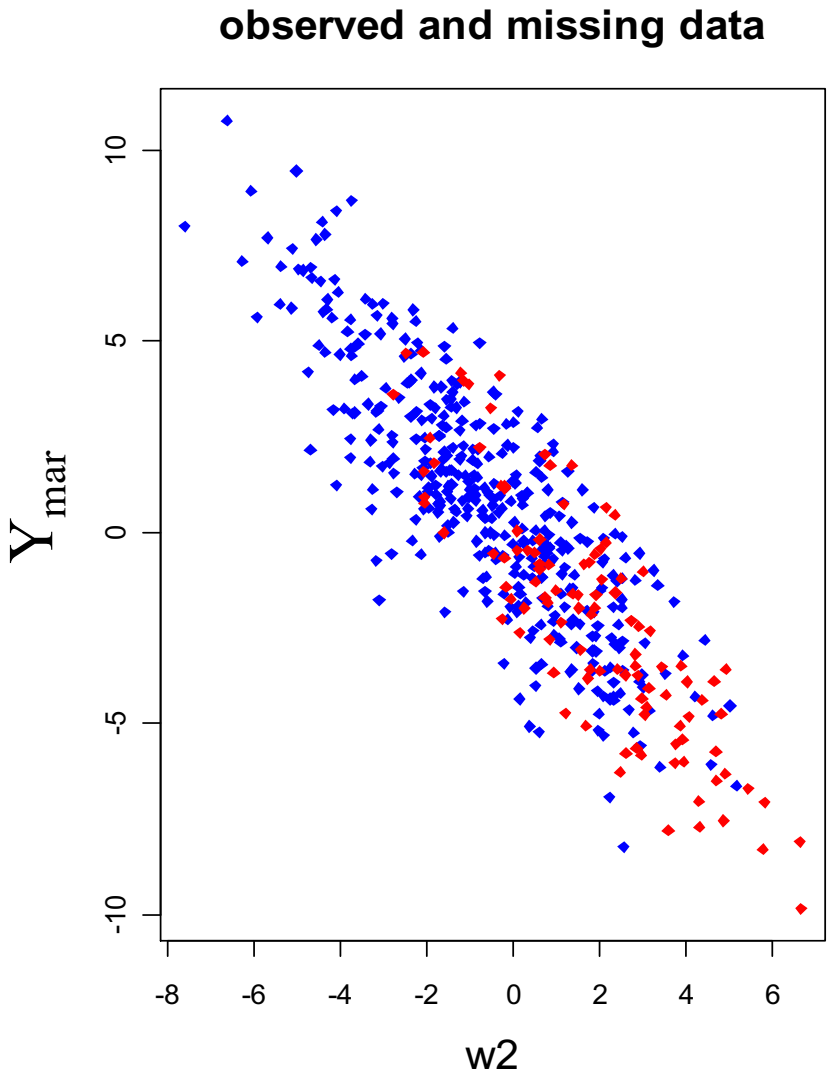
  With a long, self-administered survey, for which there is a limited amount of time for completion, fast readers will complete the survey, but slow readers will leave some questions blank at the end. However, reading speed is something that can be measured early in the questionnaire where virtually all of the respondents will provide data.

MAR data (if we could see everything!):

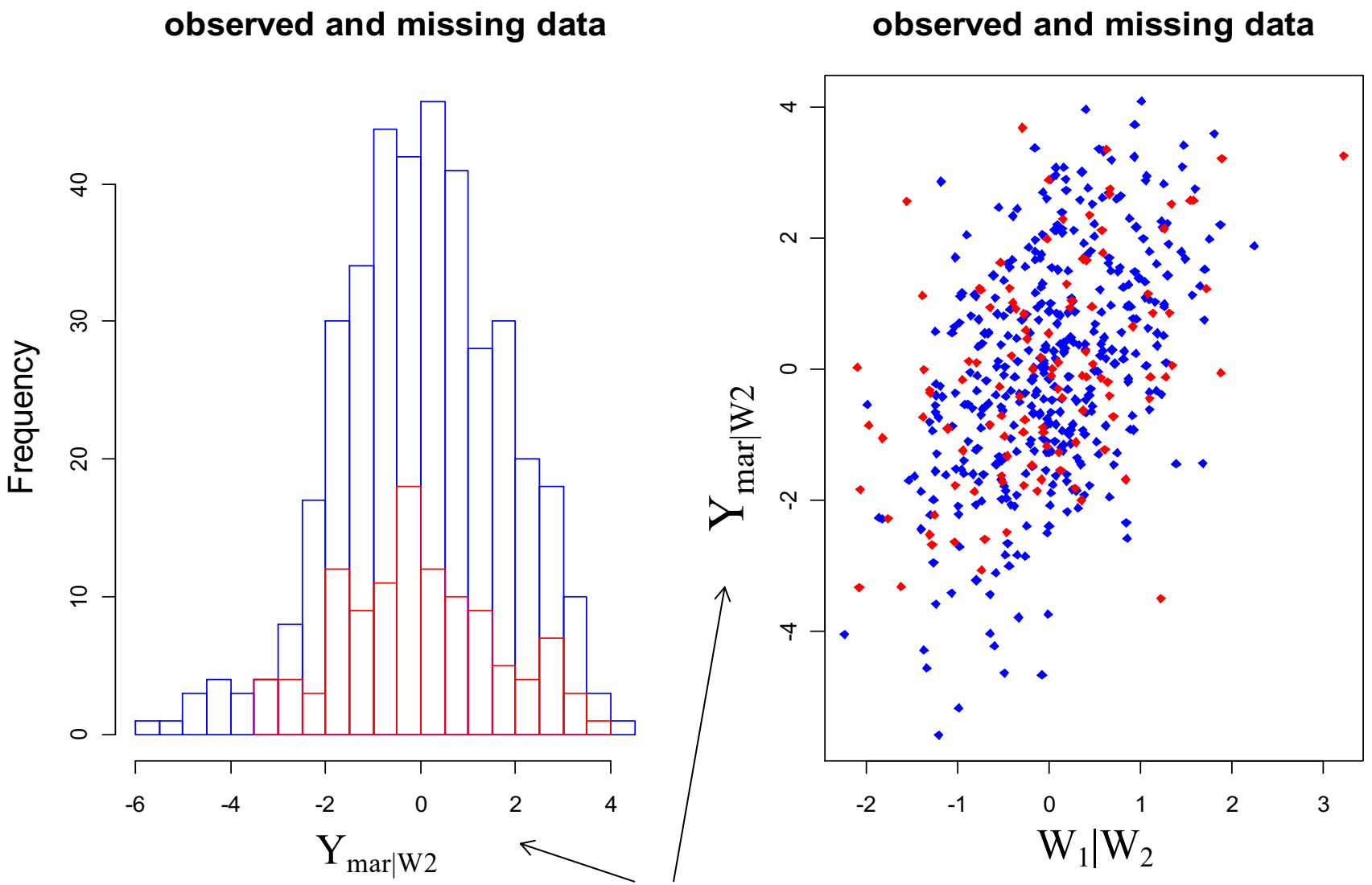


MAR data:


observed and missing data
Y_{mar}
10
5
0
-5
-10
-2
-1
0
1
2
w1


observed and missing data
Y_{mar}
10
5
0
-5
-10
-8
-6
-4
-2
0
2
4
6
w2

MAR data: Y_{mar} *conditional* on W_2



Where $Y_{mar\,|W_2}$ is the residuals from a regression of Y_{mar} on W_2

Not Missing at Random (NMAR)

- In this case missingness depends on the values of the items that are missing!
- Thus

  $P(R_1, \ldots, R_p \mid Y_1, \ldots, Y_p) \neq P(R_1, \ldots, R_p \mid Y_1^{obs}, \ldots, Y_p^{obs})$

  $P(\mathbf{R} \mid \mathbf{Y}) \neq P(\mathbf{R} \mid \mathbf{Y}^{obs})$
- One way of formalizing this is: $P(\mathbf{R} \mid \mathbf{Y}) = P(\mathbf{R} \mid \mathbf{Y}^{obs}, Z)$
- That is, NMAR missingness occurs when missingness on Y (i.e., R) is caused by *Y* itself, by some variant of Y, or by some other variable that is related to Y, but which has not been measured.
- Example: very wealthy people are less likely to report their income *and this wealth is not predicted by the other variables in the data*

Recall: Multivariate Normal Model

Assume that we have multivariate observations

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

If there are no missing values, then the likelihood is

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\theta}, \boldsymbol{\Sigma})$$

$$= \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} (\det \boldsymbol{\Sigma})^{-½} e^{-½(\boldsymbol{x}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})}$$

$$= (2\pi)^{-\frac{np}{2}} (\det \boldsymbol{\Sigma})^{-\frac{n}{2}} e^{-½ \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})' \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})}$$

Q: How do we compute $f(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\Sigma}) =$

$(2\pi)^{-\frac{p}{2}} (\det \boldsymbol{\Sigma})^{-½} e^{-½(\boldsymbol{x}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})}$ when the $\boldsymbol{x}_i$ is missing?

A: Impute a value.

Note: The worst idea is to impute the average!

Likelihood with missing data

Assume MAR. Then the likelihood of observed data for subject *i* is

$$f\left(\boldsymbol{r_i}, \{x_{ij}: r_{ij} = 1\} \middle| \boldsymbol{\theta}, \boldsymbol{\Sigma}\right) = p(\boldsymbol{r_i}) \times f\left(\{x_{ij}: r_{ij} = 1\} \middle| \boldsymbol{\theta}, \boldsymbol{\Sigma}\right)$$

$$= p(\boldsymbol{r_i}) \times \int \left[f(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\Sigma}) \prod_{J: r_{ij}=0} dx_{ij} \right]$$

That is, we multiply $p(\boldsymbol{r_i})$ with the marginal distribution of the observed variables, after integrating out the missing variables.
For example, if $\boldsymbol{x}_i = (x_{i1}, NA, NA, x_{i4})$, then $\boldsymbol{r_i} = (1, 0, 0, 1)$ and

$$f(\boldsymbol{r}_i, x_{i1}, x_{i4} | \boldsymbol{\theta}, \boldsymbol{\Sigma}) = p(\boldsymbol{r_i}) \times \iint f(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\Sigma}) dx_{i2} dx_{i3}$$

So, the correct thing to do when data are missing at random is to integrate over the missing data to obtain the marginal probability of the observed data.

Algorithm

- Let $\boldsymbol{X}$ be the $n \times p$ matrix of all data, both observed and missing.
- Let $\boldsymbol{R}$ be the $n \times p$ missing pattern matrix as defined before.
- The matrix $\boldsymbol{X}$ consists of two parts:
- $\boldsymbol{X}_{\text{obs}} = \{x_{ij}: r_{ij} = 1\}$ is the observed data.
- $\boldsymbol{X}_{\text{mis}} = \{x_{ij}: r_{ij} = 0\}$ is the unobserved or missing data.
- Note that $\boldsymbol{X}_{\text{mis}}$ has to be treated as unknown parameter!
- Goal: Obtain samples from the posterior distribution

$$f(\boldsymbol{\theta}.\boldsymbol{\Sigma}, \boldsymbol{X}_{mis} | \boldsymbol{X}_{obs})$$

- Solution: Run a three-stage Gibbs sample that iterates between

$$f(\boldsymbol{\theta} | \boldsymbol{X}_{obs}, \boldsymbol{\Sigma}, \boldsymbol{X}_{mis})$$
$$f(\boldsymbol{\Sigma} | \boldsymbol{X}_{obs}, \boldsymbol{\theta}, \boldsymbol{X}_{mis})$$
$$f(\boldsymbol{X}_{mis} | \boldsymbol{X}_{obs}, \boldsymbol{\theta}, \boldsymbol{\Sigma})$$