Missing Data

Dobrin Marchev

January 29, 2021

Getting started with `R` and `RStudio`

Installing R and/or RStudio

Google the following:

R

or go to https://cran.r-project.org/

and follow the links to download and install R

then go to

https://www.rstudio.com/

and download and install

What is R/RStudio?

- R is software for graphing, manipulating and analyzing data;
- It consists of a console, source and graph windows;
- R is an expression language;
- It is case sensitive!
- Each command is either an expression or an assignment;
- Commands are separated by ‘;’ or new line;
- Commands are grouped together with ‘{ }’;
- Comments start with ‘#’;
- Vertical arrow recalls the previous commands;
- Each R entity is an object, like vector, matrix or data frame.

Some basic R examples:

- Create a vector of numbers: `x = c(4, 3, 5, 6, 20)`
- Manipulate a vector: `x^2; 1/x; x + 3; 2*x; ...`
- Or use built-in functions: `sqrt(x); sum(x); length(x);`
- Logical vectors: `x < 20`
- Getting specific elements: `x[2]; x[-2]; x[2:4];x[x<20]`
- Creating matrices: `A = matrix(1:6, 3, 2)`
- Multiplying vectors and matrices: `y = c(1,2); A%*%y`
- Combining vectors into a matrix: `z = c(3,4); cbind(y,z)`
- Data are stored into data frames: `head(ChickWeight)`
- Make names of columns visible: `attach(ChickWeight)`

R packages:

- See all packages: `library()`
- In Rstudio simply click on the “Packages” tab
- To install a new package: `install.packages("name")`
- Usually automatically installs dependencies
- Note: this just saves the package to the hard disk!
- To load a package into the memory: `library(name)`
- Note: no quotes for library, but quotes for installation
- In Rstudio you can simply checkmark the package
- To see new datasets available with pacage: `data()`
- To know what’s in a package, read package help file.

Why should we care about missing data?

Simple example in R:

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

Why should you care about missing data?

- Improperly dealing with missing data can cause:
  - Bias
  - Efficiency loss (particularly within subgroups)
  - Incorrect standard errors
- You may also be interested in the missing data mechanism (*why* are these people missing?)

Example: Multiple regression

- The standard approach to deal with missing data in multiple regression is to delete the entire row if any of the variables have a missing value.
- Suppose we have a sample of 1000 observations (rows)
- There are 20 predictor variables
- Each variable has 5% probability to have missing data
- Assume missing data on one variable is independent of the missing data on the other variables
- What is the expected number of complete cases? (That is, cases for which all 20 variables are observed.)

Example: Multiple regression

- Answer: $1000*(0.95^{20})$ or 358 complete cases!
- If probability of missing data = 10%, then we will have only about 122 complete cases!!!

- Questions to ask ourselves:
  - Can we salvage some info from the remaining 642 cases? That is, are we making the best use of our data?
  - How do we perform best subset selection if each variable has different sample size? For example, are the R^2 comparable?
  - Do the estimated coefficients generalize to the population?
  - Do we have enough cases?
  - If we compare two samples, are the differences due to the model or to the missing data?

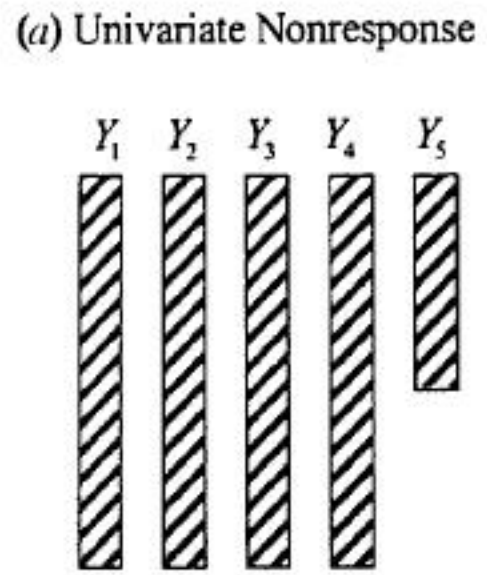
Outline of the Course

1) **Missing Data Mechanisms. How are missing data generated and why should we care? Complete Case Analysis. Getting comfortable with R.**
2) **Simple missing data fixes: available case, LVCF, mean imputation, dummy variable methods. Getting even more comfortable with R**
3) **More complicated missing data fixes: weighting, hotdecking, regression imputation**
4) **Building blocks and overview of multiple imputation (including regression imputation with noise)**
5) **Multiple imputation in practice (software in R, simple analyses, and diagnostics)**
6) **Multiple imputation in practice (more complicated models and considerations, more advanced diagnostics)**
7) **More advanced imputation and other missing data methods**

Warnings

- Does this course discuss all missing data approaches?
- No!
- There are many ways of addressing missing data problems. However many of them are specific to the given analysis model. This course covers more general methods that can be used in conjunction with virtually any analysis method.

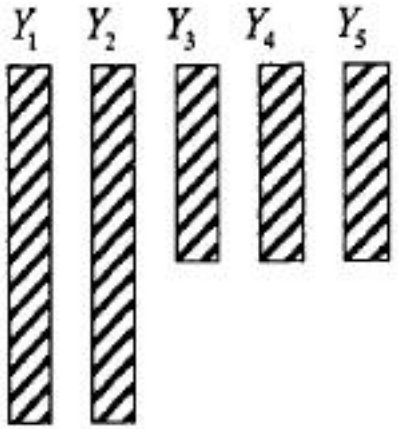
Examining Patterns of Missing Data



- Missingness is confined to a single variable
- For example, Y_5 is yield of crop
- and $Y_1, \ldots, Y_4$ are variety, type of fertilizer, temperature, intended to be fully observered
- Y_5 could be missing because of lack of germination or incorrectly recorded data

Examining Patterns of Missing Data

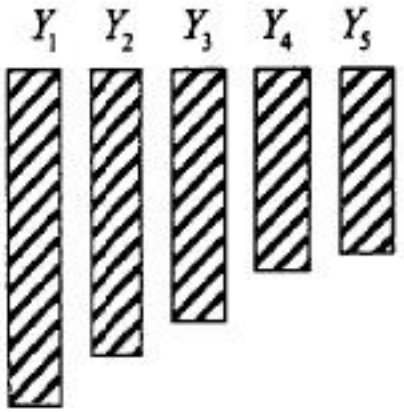
(*b*) Multivariate Two Patterns



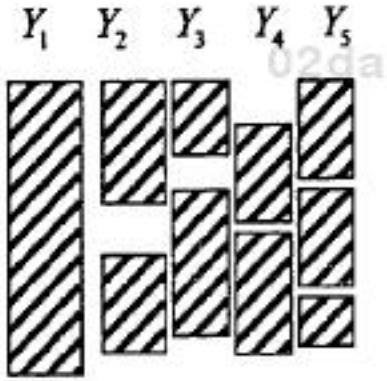
- A set of variables all missing on the same set of cases
- For example, unit nonresponse in surveys
- Questionnaire is administered: subset of sample individuals do not complete the survey because of noncontact, refusal, or other reason

Examining Patterns of Missing Data

(*c*) Monotone



(*d*) General



Ex: Longitudinal studies

Notation

- Sample data matrix is usually denoted $\boldsymbol{Y}$, 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 $\boldsymbol{Y}$ 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, $Y_1, \ldots, Y_p$, that indicate whether a given value of the corresponding Y variable is observed (= 1) or missing (= 0)

Y_1	Y_2	Y_3	Y_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

Creating the missing data matrix in R

```
# First create 4x3 data matrix
Y = matrix(1:12, 4, 3)
# Make some missing elements
Y[2, 3] = Y[3, 2] = Y[3, 3] = NA

# is.na function returns TRUE or FALSE, depending if missing or observed
is.na(Y)
# ! negates a statement
!is.na(Y)
# Converting TRUE FALSE to 1 0:
R = (!is.na(Y))*1
```

Missing Data Mechanisms (MDMs)

- Rubin's (1976) missing data theory involves two sets of parameters: the parameters that address the research questions (i.e., the parameters that you would have estimated had there been no missing data) and the parameters that describe the probability of missing data (usually denoted φ).
- Researchers rarely know why the data are missing, so it is impossible to describe φ with any certainty.
- Rubin's work is important because he clarified the conditions that need to exist in order to accurately estimate the substantive parameters without also knowing the parameters of the missing data distribution (i.e., φ).

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)

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/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

Missing Completely at Random

- $P(R_1, R_2, \ldots, R_p \mid Y_1, Y_2, \ldots, Y_p) = P(R_1, R_2, \ldots, R_p)$
- More formally:

$$P(\boldsymbol{R} = \boldsymbol{0} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \varphi) = P(\boldsymbol{R} = \boldsymbol{0} \mid \varphi)$$

- $\boldsymbol{R} \perp \boldsymbol{Y}$ ($\boldsymbol{R}$ and $\boldsymbol{Y}$ are independent)
- This means that whether any given value is missing is *completely random* (doesn't depend on anything).
- This is generally not a plausible assumption. Usually, certain types of people are much more likely than others to have missing data.
- When could this happen?
  - A bunch of records are lost
  - Missing by design

Let's simplify

- What if we just have two variables:
- W_1 completely observed
- Y with some missing values
- Implicitly of course we have a third, R
- Y is missing completely at random, that is

  $R \perp Y, W_1$
- Therefore, the missing values should look like a random subsample of the full sample (as should the observed values)
- This also implies that the joint distribution of Y and W_1 should be the same whether $R = 0$ or $R = 1$ (i.e. whether Y is observed or not)

Simulated data

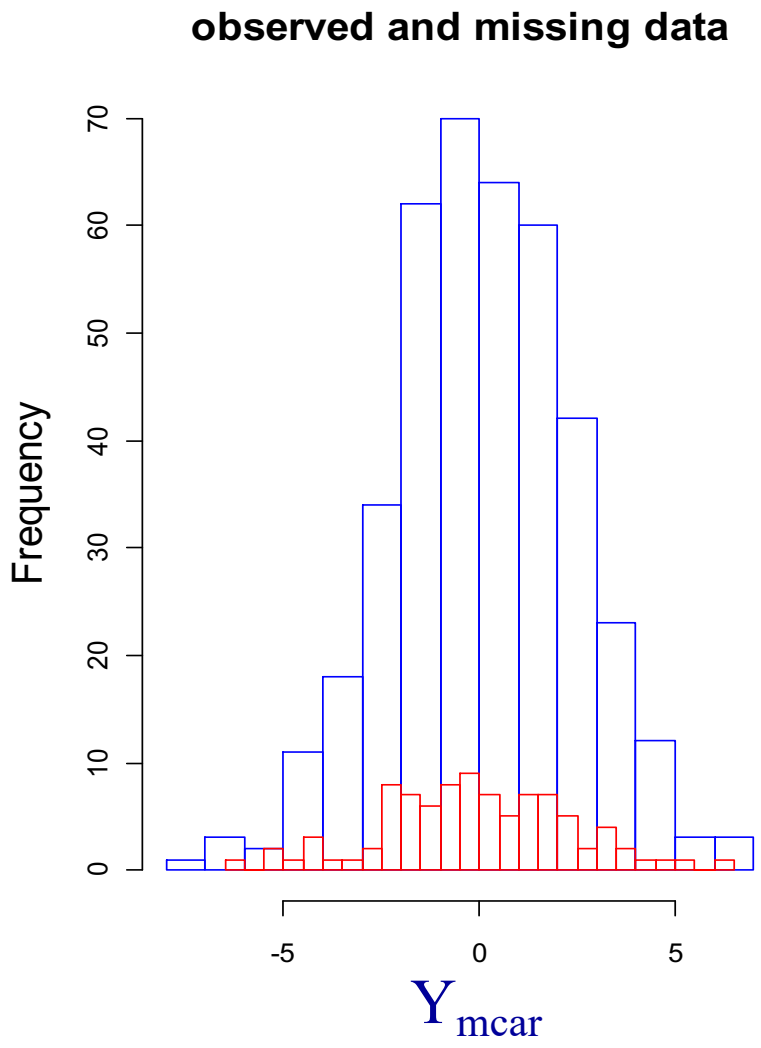
I've simulated some data to graphically illustrate some of the properties of these concepts (how to simulate more complicated missing data will be explained next lecture, but in the R file for today is the MCAR case).

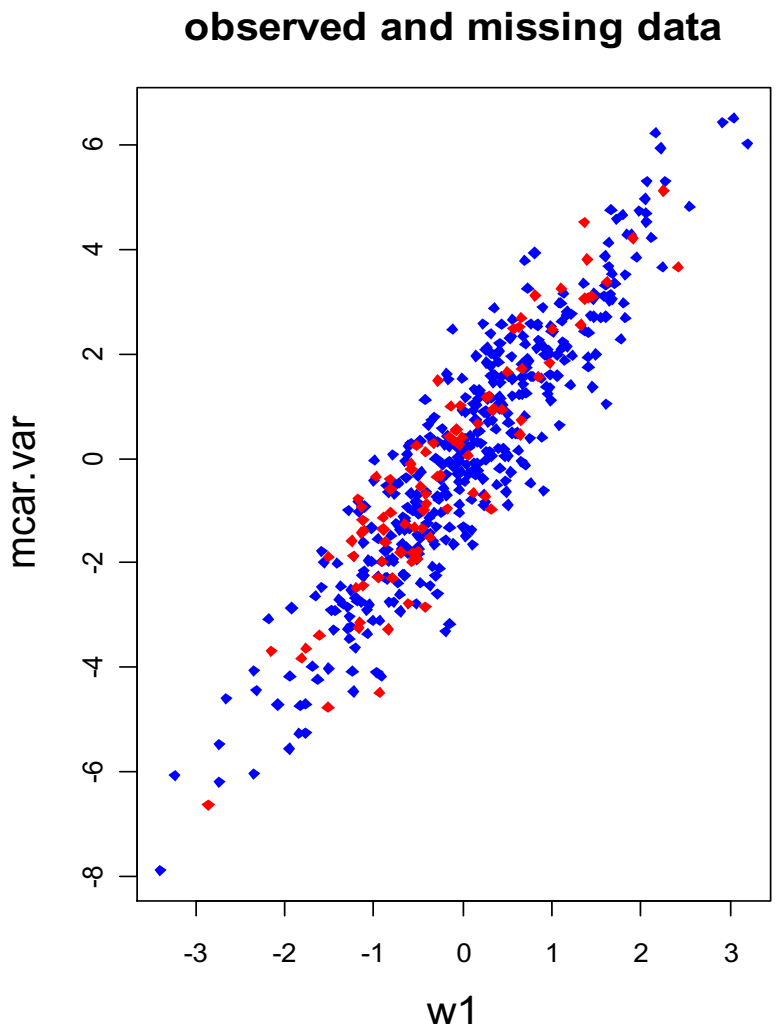
- W_1 and W_2 are fully observed variables
- Y_{mcar} is a variable with MCAR missing data
- Y_{mar} is a variable with MAR missing data
- Y_{nmar} is a variable with NMAR missing data

MCAR data (if we could observe it!):

blue = observed

red = missing





Testing for MCAR?

- Is it possible to test whether or not your data are MCAR?
- Well you can never say for sure that your data are MCAR…. however you may be able to show that they are (likely) not…
- We'll discuss more in a bit

Missing at Random (MAR)

- $P(R_1, \ldots, R_p \mid Y_1, \ldots, Y_p) = P(R_1, \ldots, R_p \mid Y_1^{obs}, \ldots, Y_p^{obs})$
- $P(\boldsymbol{R} \mid \boldsymbol{Y}) = P(\boldsymbol{R} \mid \boldsymbol{Y}^{obs})$
- Here missingness depends on observed values of the variables.
- A simple version of this is $P(R_1 \mid Y_1, \boldsymbol{W}) = P(R_1 \mid W)$, where $\boldsymbol{W}$ is a subset of fully observed variables in the matrix $\boldsymbol{Y}$
- 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.

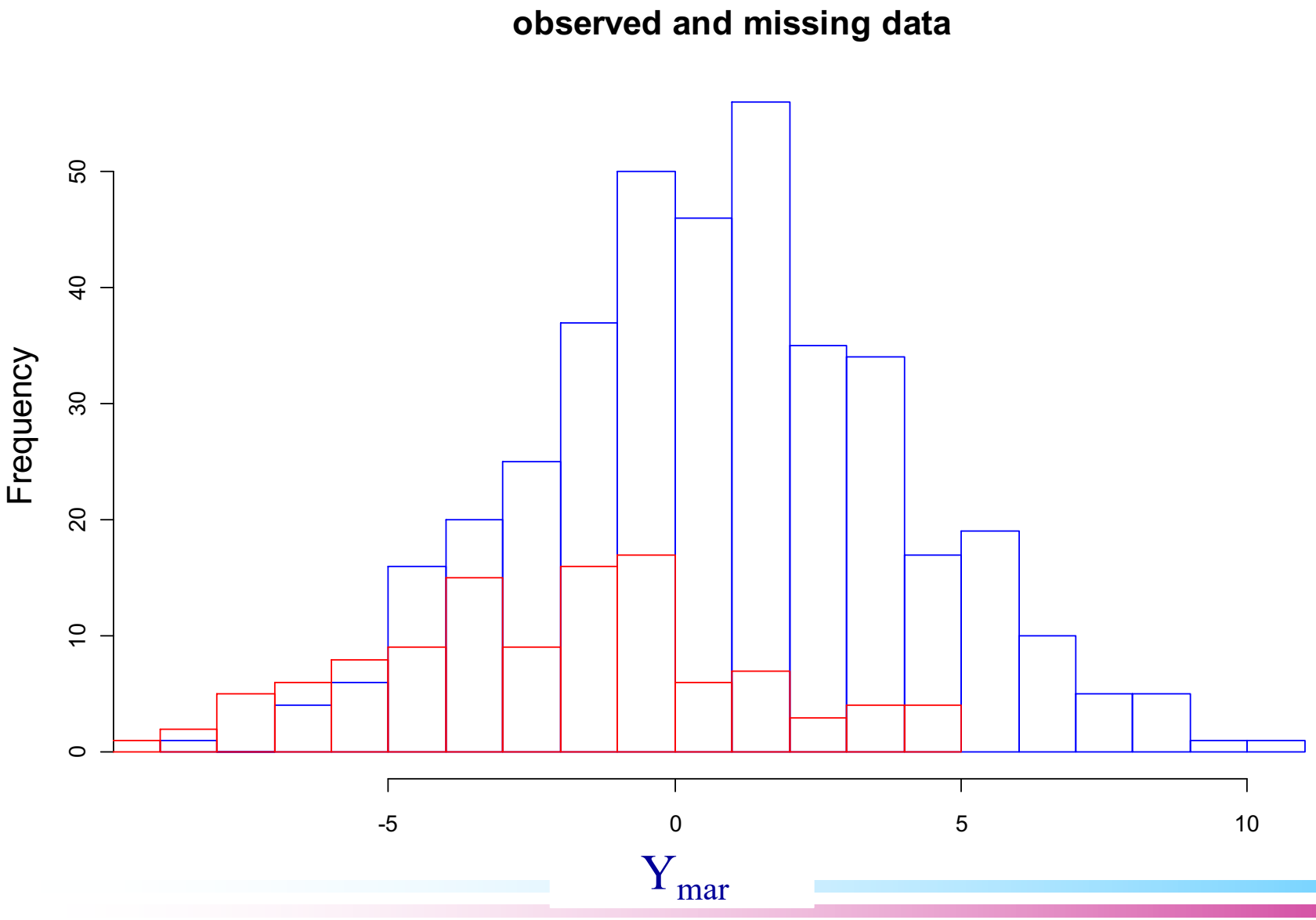
Let's simplify

- What if we just have three variables, W_1, W_2, and Y
- Implicitly, of course, we have a fourth, R
- We can formalize the property that the missingness for Y is missing at random by saying that

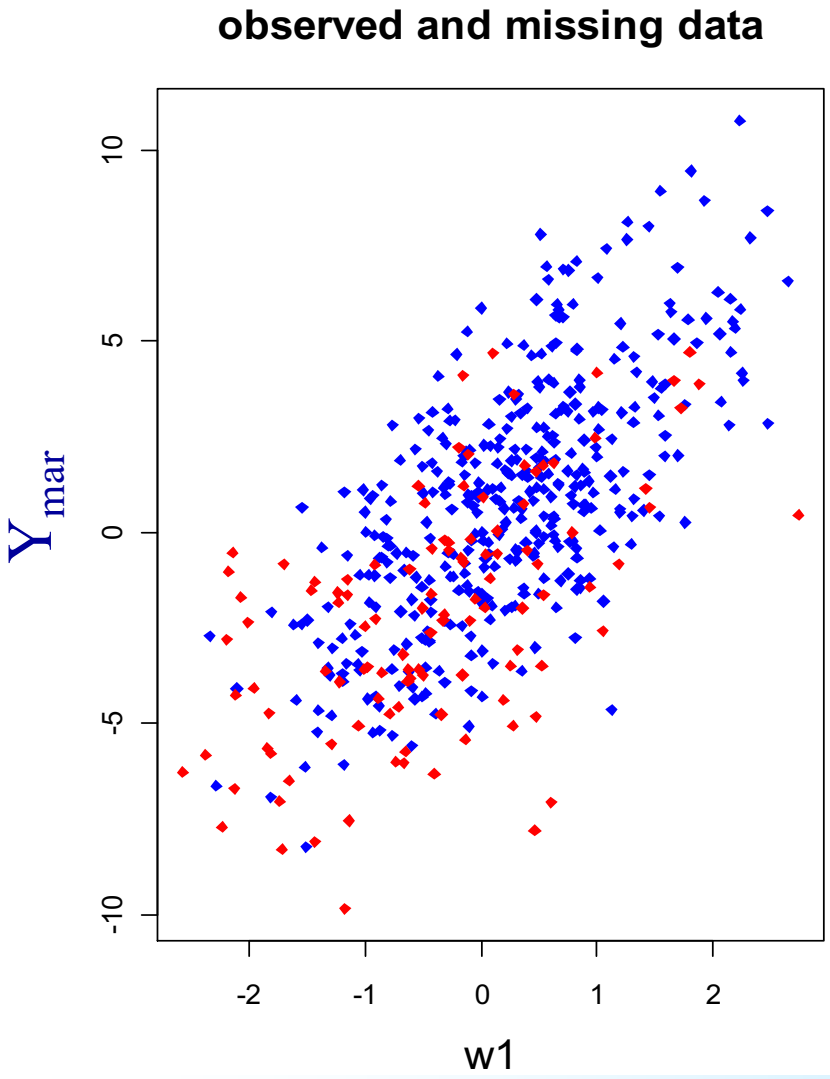
  $R \perp Y \mid W_1, W_2$
- In the example on next page it's actually true that

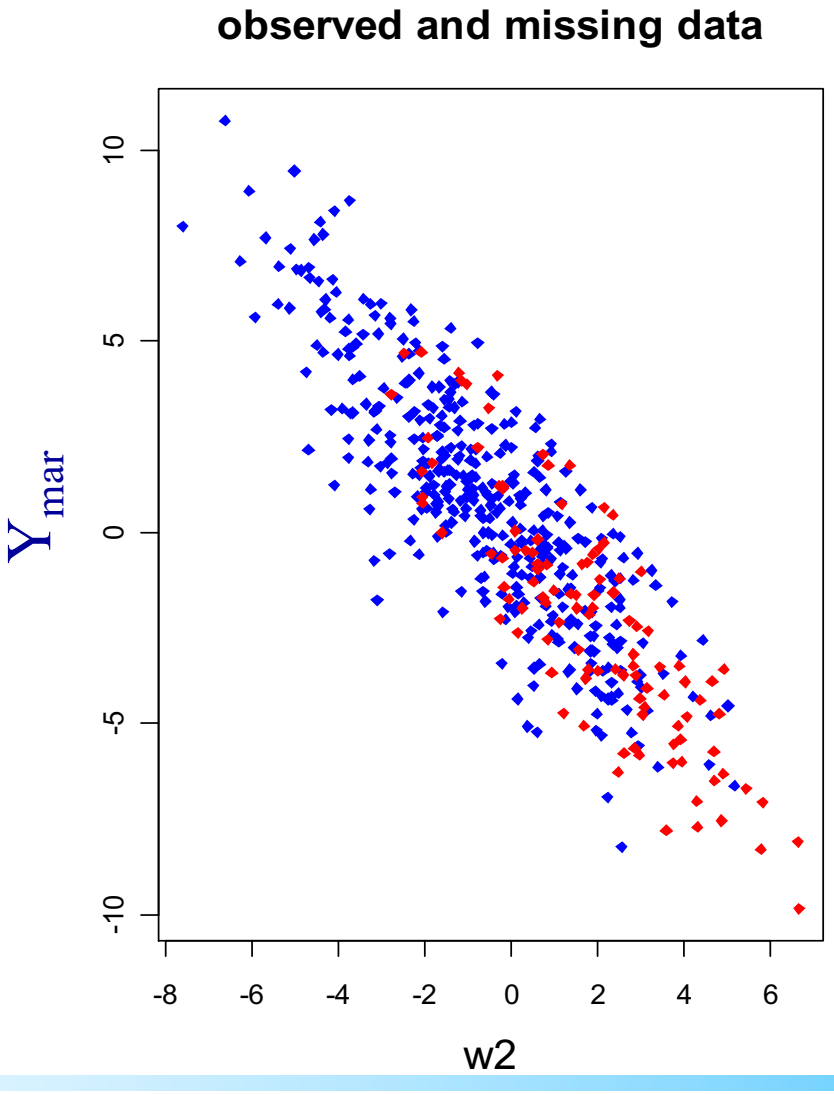
  $R \perp Y \mid W_2$
- Therefore the missing values should look like a random sample of the full sample (as should the observed values) within subsets of the data that have the same values of W_2
- This also implies that the joint distribution of Y and W_1 should be the same whether Y is observed or not *after conditioning on* W_2

MAR data (if we could see everything!):

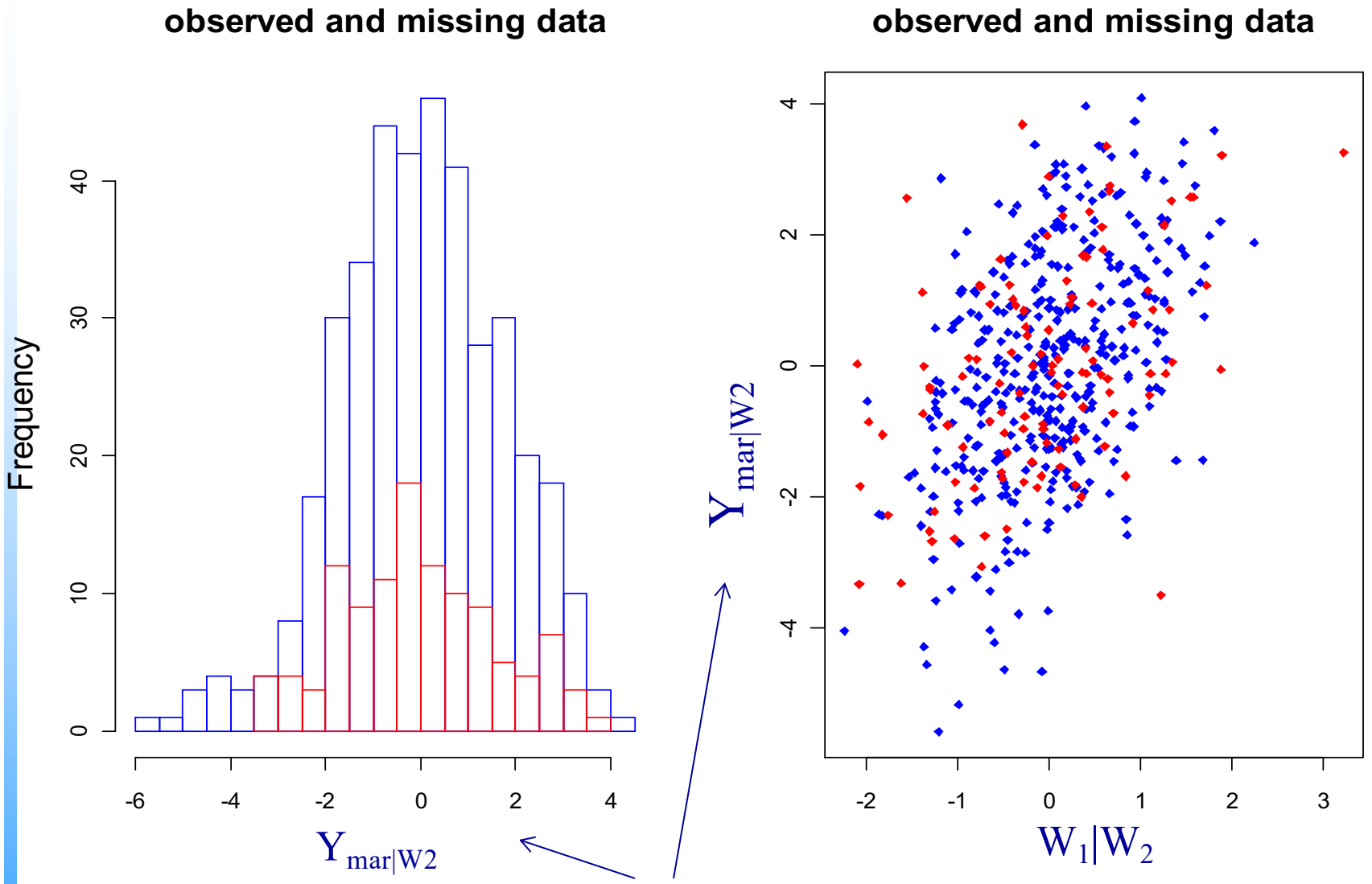


MAR data:


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

MAR: creating MAR

- Missingness on *Y* is a linear function of *Z*:

  if Z = 1, the probability that *Y* is missing, P(R = 0) = 0.20

  if Z = 2, the probability that *Y* is missing, P(R = 0) = 0.40

  if Z = 3, the probability that *Y* is missing, P(R = 0) = 0.60

  if Z = 4, the probability that *Y* is missing, P(R = 0) = 0.80

- A more complicated form of missing at random might depend on another variable *X* in the model. Divide the data into clusters and for each calculate the correlation between X and Z. Then:

  if corr(X, Z) = high, prob. that Y is missing, P(R = 0) = 0.80

  if corr(X, Z) = low, prob. that Y is missing, P(R = 0) = 0.20

MAR: adding back a bit of complexity

- What if we have four variables in our dataset: W_1, W_2, Y_{mar}, and Y_{mcar}?
- A more complicated form of missing at random might look like the following

  $R_{mar.var} \perp Y_{mar} \mid W_1, W_2, Y_{mcar}^{obs}$
- What does this mean??
- In practice this could mean that

  $R_{mar.var} \perp Y_{mar} \mid W_1, W_2, Y_{mcar}, R_{mcar} = 1$

  and

  $R_{mar.var} \perp Y_{mar} \mid W_1, W_2, R_{mcar} = 0$

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*

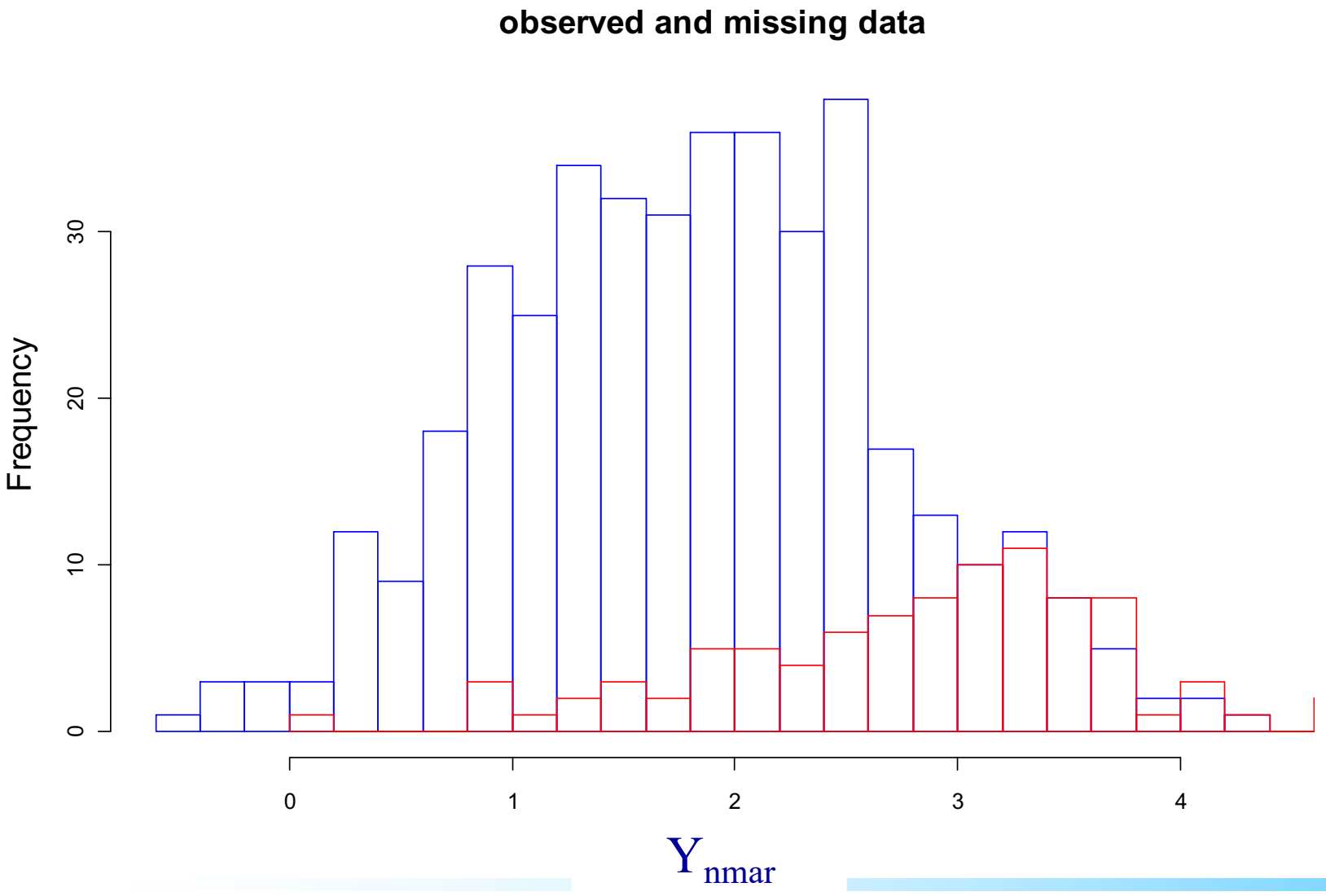
NMAR

- With our simulated data if there were only nmar.var Y, W_1, and W_2, this would mean that

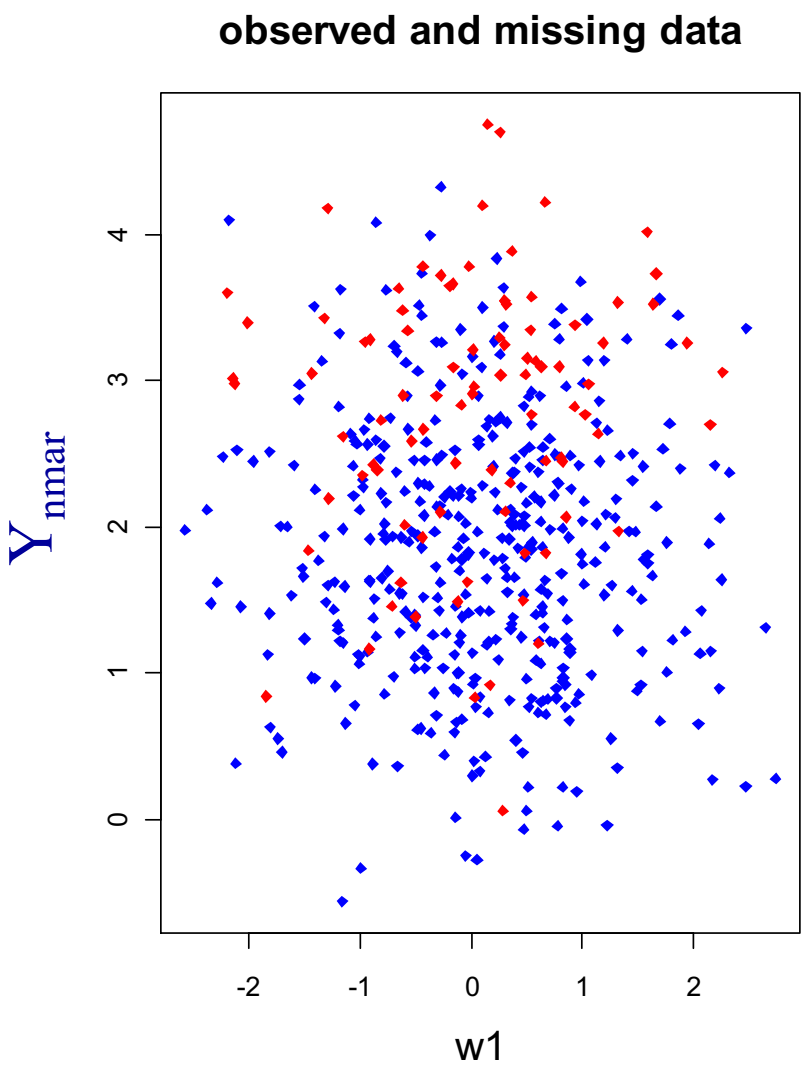
  $R_{nmar} \perp\!\!\!\not\perp Y_{nmar} \mid W_1, W_2$
- If all the variables were in the dataset it would also mean that

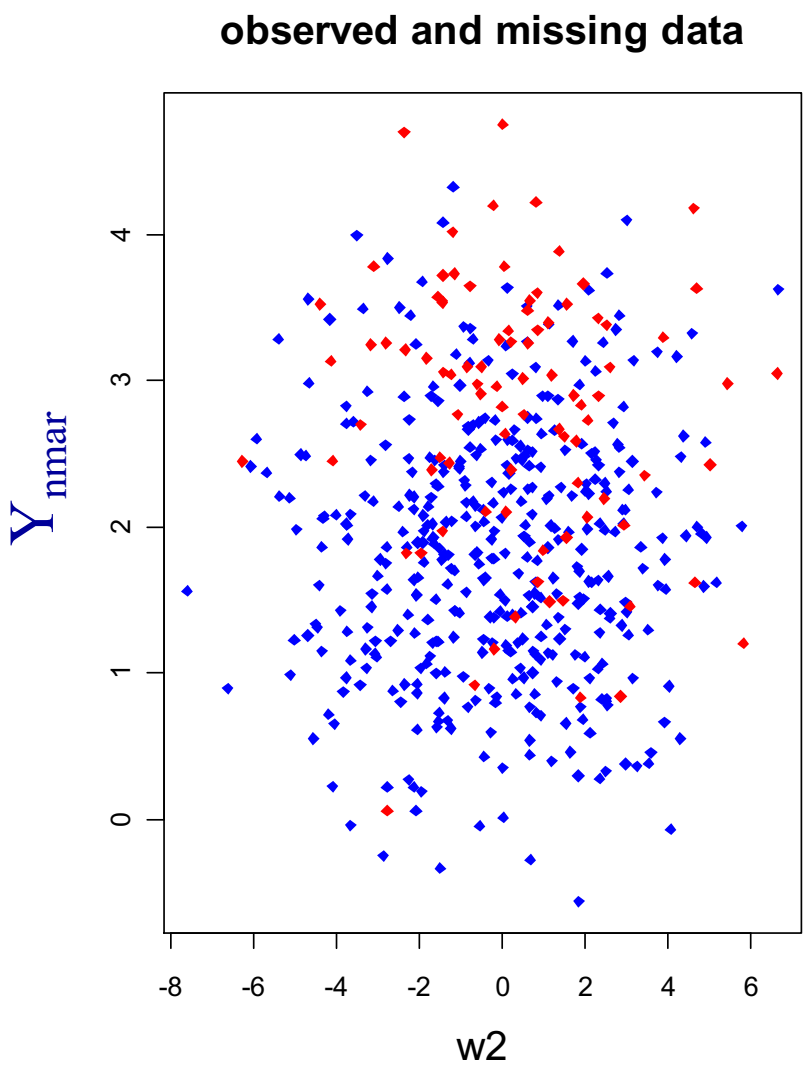
  $R_{nmar} \perp\!\!\!\not\perp Y_{nmar} \mid W_1, W_2, Y_{mar}^{obs}, Y_{mcar}^{obs}$
- In words, this means that even after conditioning on the other variables the distribution of Y_{nmar} will be different between those with missing data and those without

NMAR data (if we could observe it):

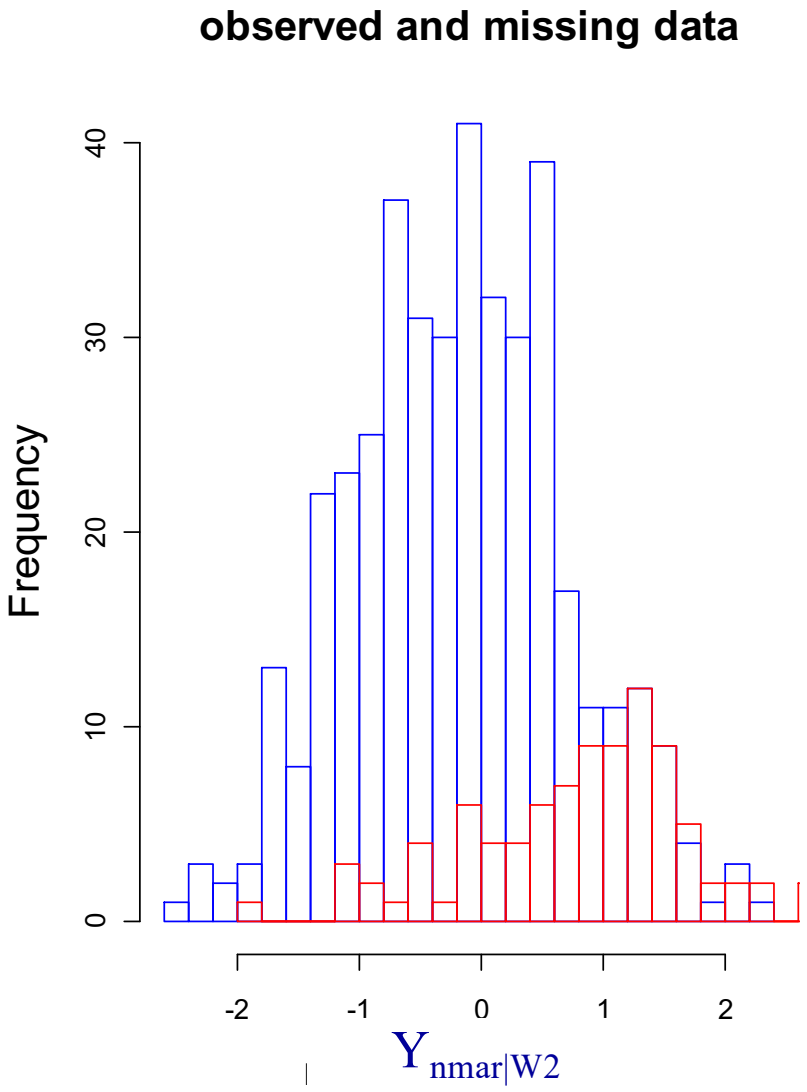


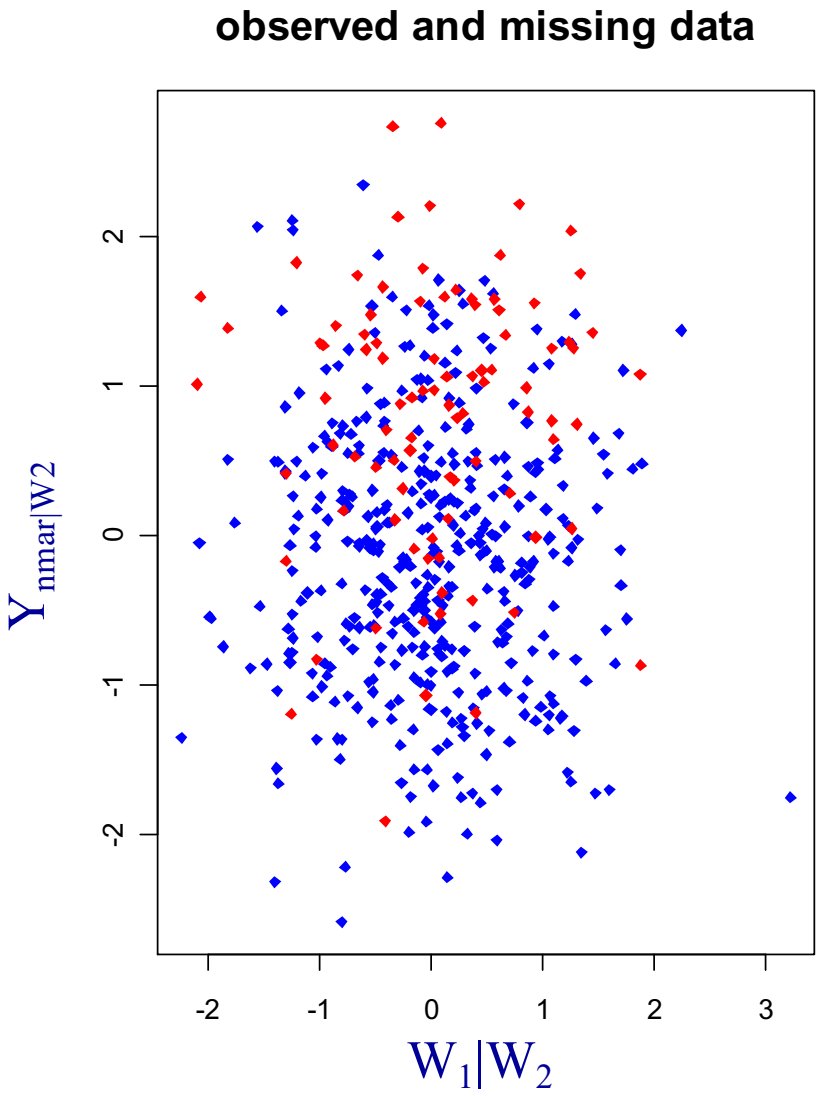
NMAR data:


observed and missing data
Y nmar
0
1
2
3
4
-2
-1
0
1
2
w1


observed and missing data
Y nmar
0
1
2
3
4
-8
-6
-4
-2
0
2
4
6
w2

NMAR data: nmar.var, *conditional* on W_2





Ignorable missing data mechanisms

- MCAR and MAR are both *ignorable* missing data mechanisms
- There are some technical details about the distinction between these terms that we'll ignore (the parameters of the missing data mechanism must be distinct from those of the model for the data)
- The term ignorable reflects the fact that for these missing data mechanisms we can make inferences using our data without having to include a model for the missing data mechanism within our analysis model
- NMAR is a non-ignorable missing data mechanism
- MAR is ignorable in the sense it is sufficient to include the variable *Z*, but not to know the exact model that generates the missingness.

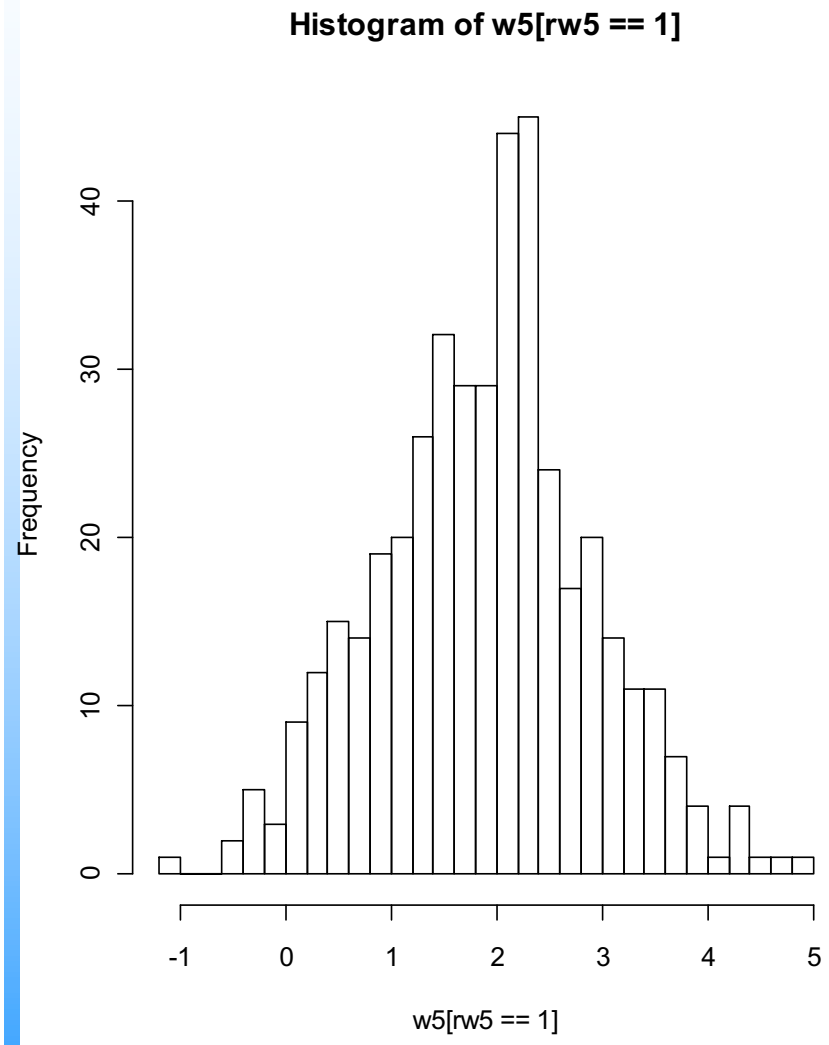
What can we determine about missingness mechanisms by looking at the data at hand?

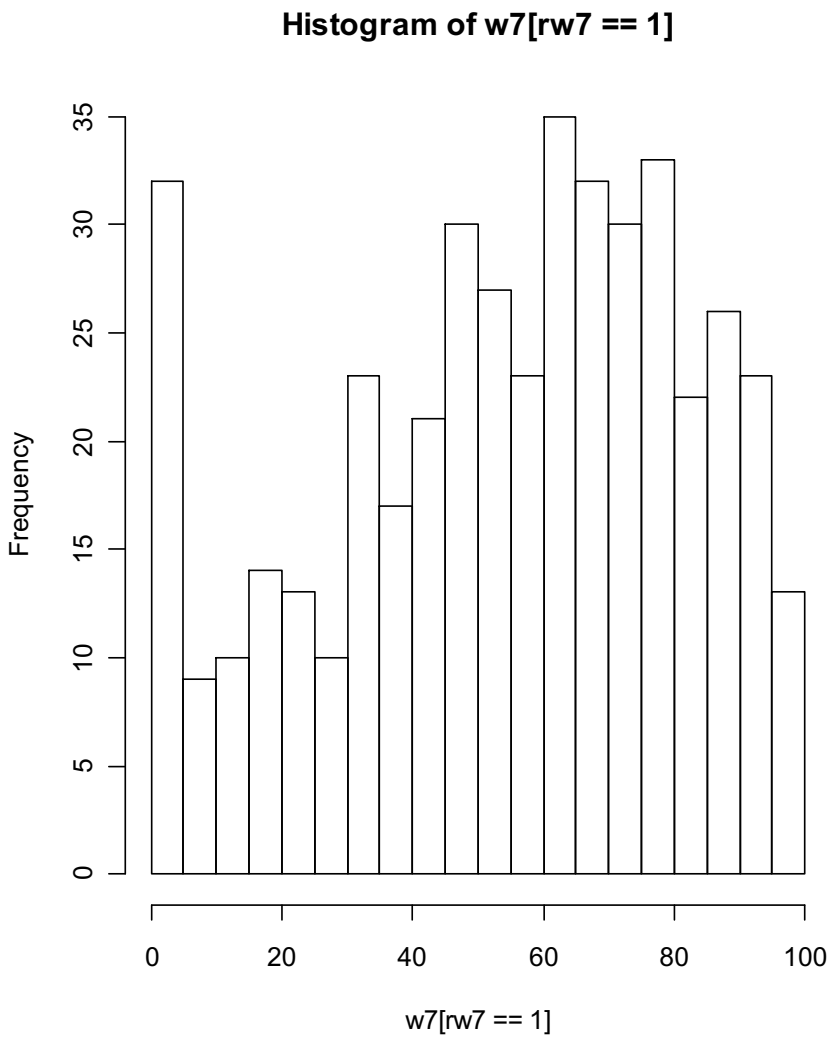
- The conventional wisdom is that one cannot know whether the missingness is MAR or NMAR ☹
- When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption.
- In general, there is no way to test whether MAR holds in a dataset, except by obtaining followup data from nonrespondents ☹

What can we learn about what kind of missing data mechanism we have, using just our observed data?

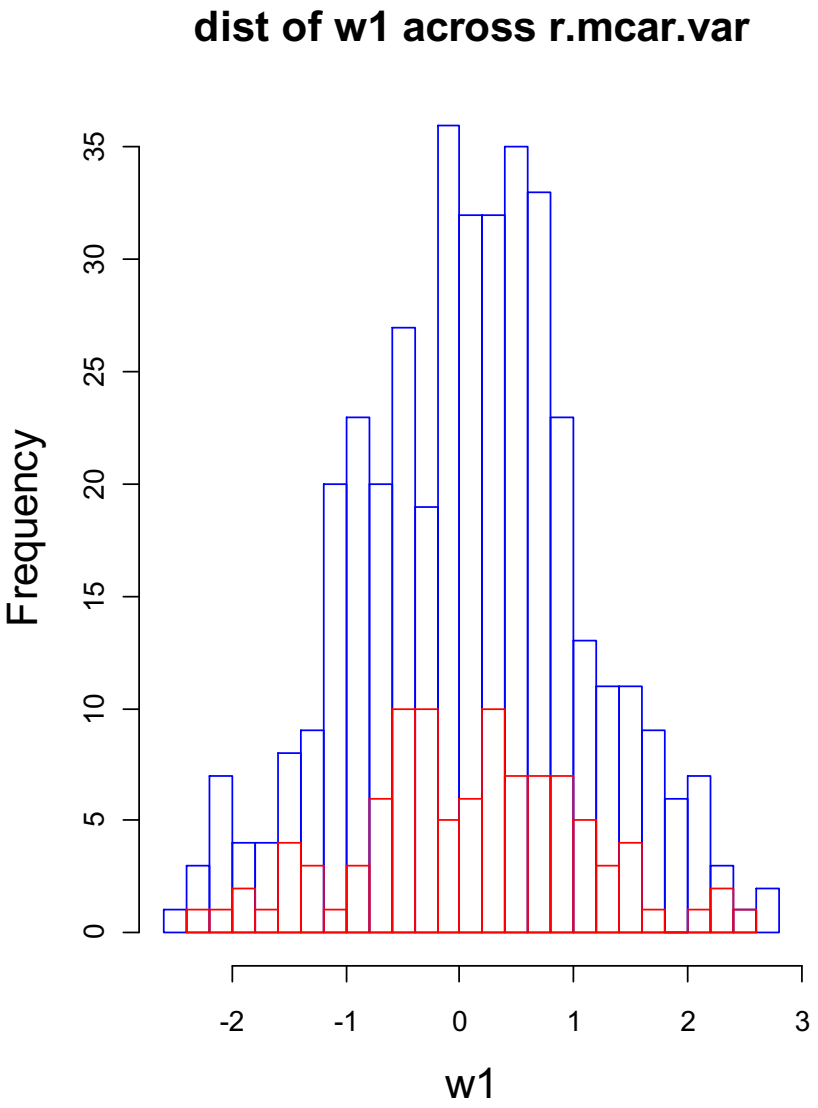
- In extreme situations we may be able to get information just from simple univariate plots
- We can look at the distribution of our fully observed data across levels of missingness of the other variables using
  - plots
  - tables that compare means or other summary statistics
- We can also do this using distributions of partially observed data but then the waters start to get muddied
- (We'll explore other diagnostics once we start using multiple imputation)

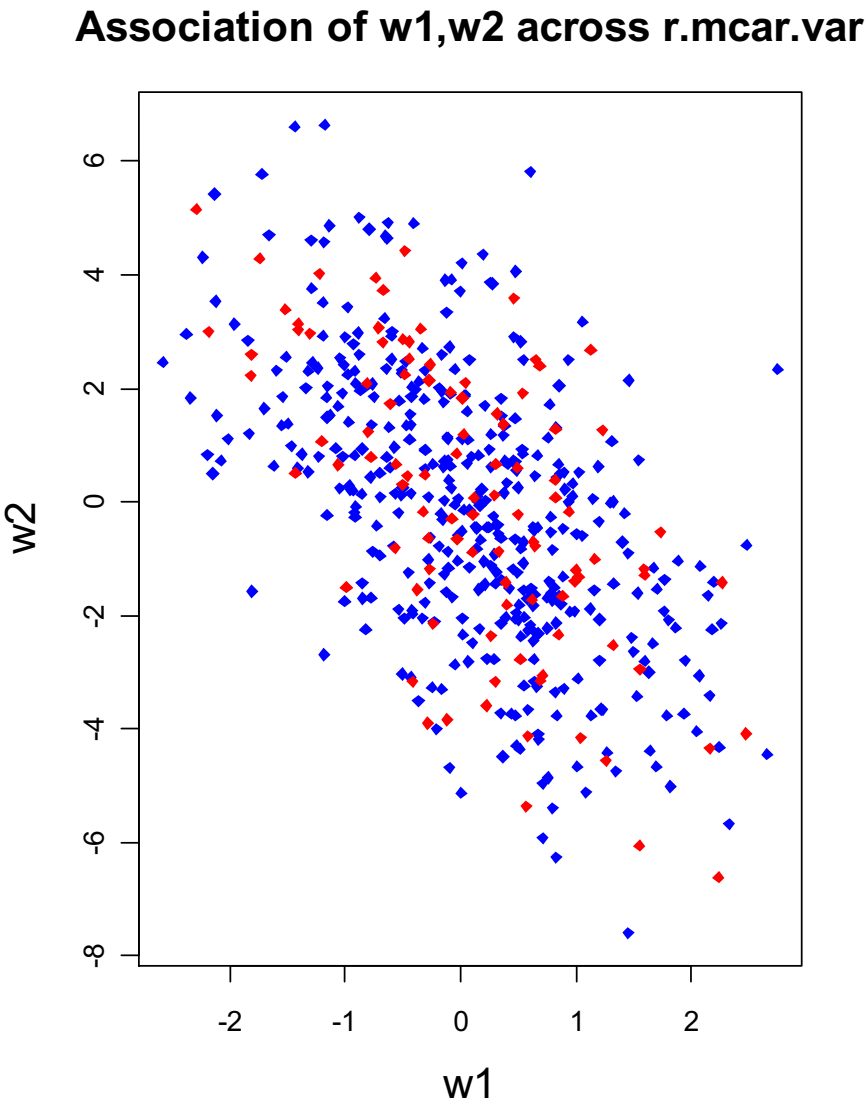
Univariate checks



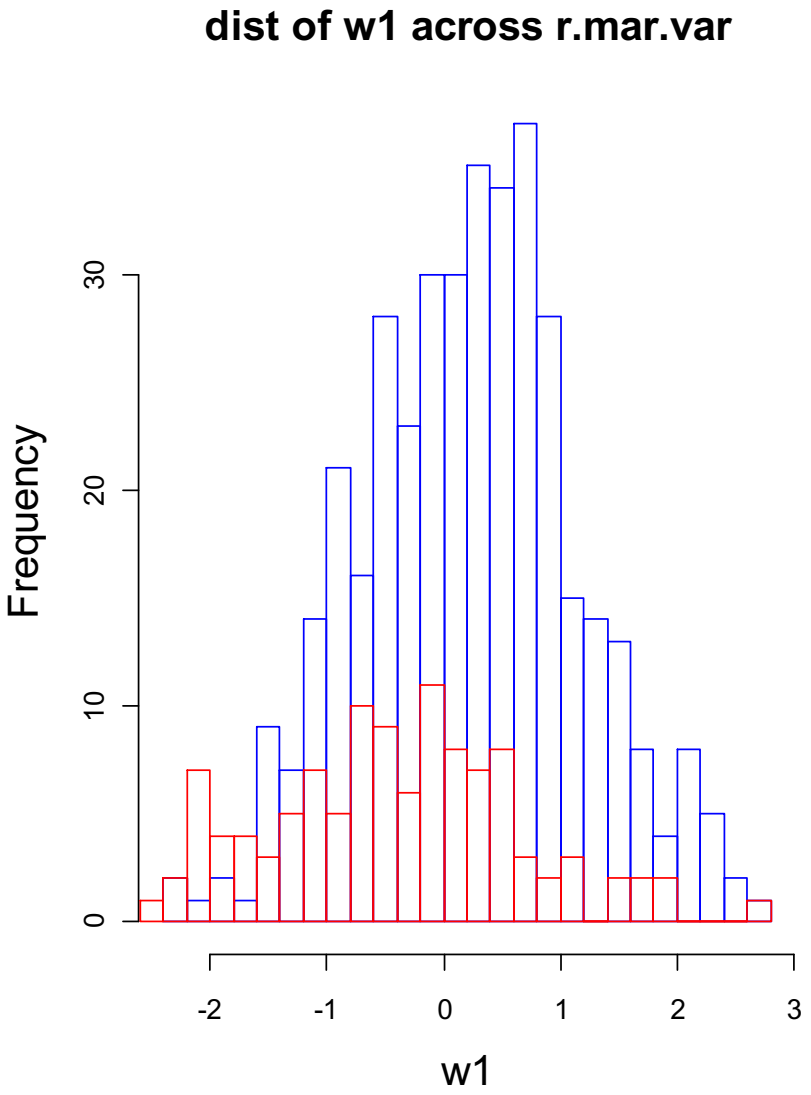


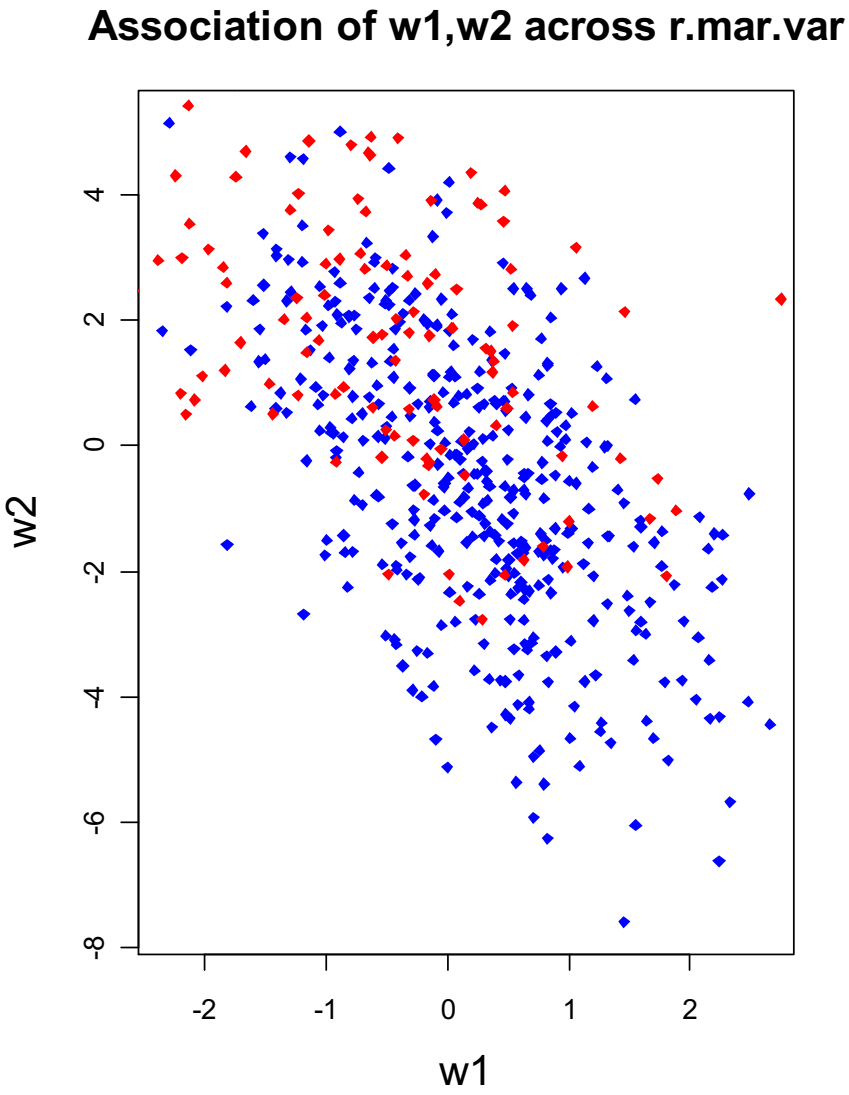
Checks regarding missingness of Y_{mcar} by using observed vars



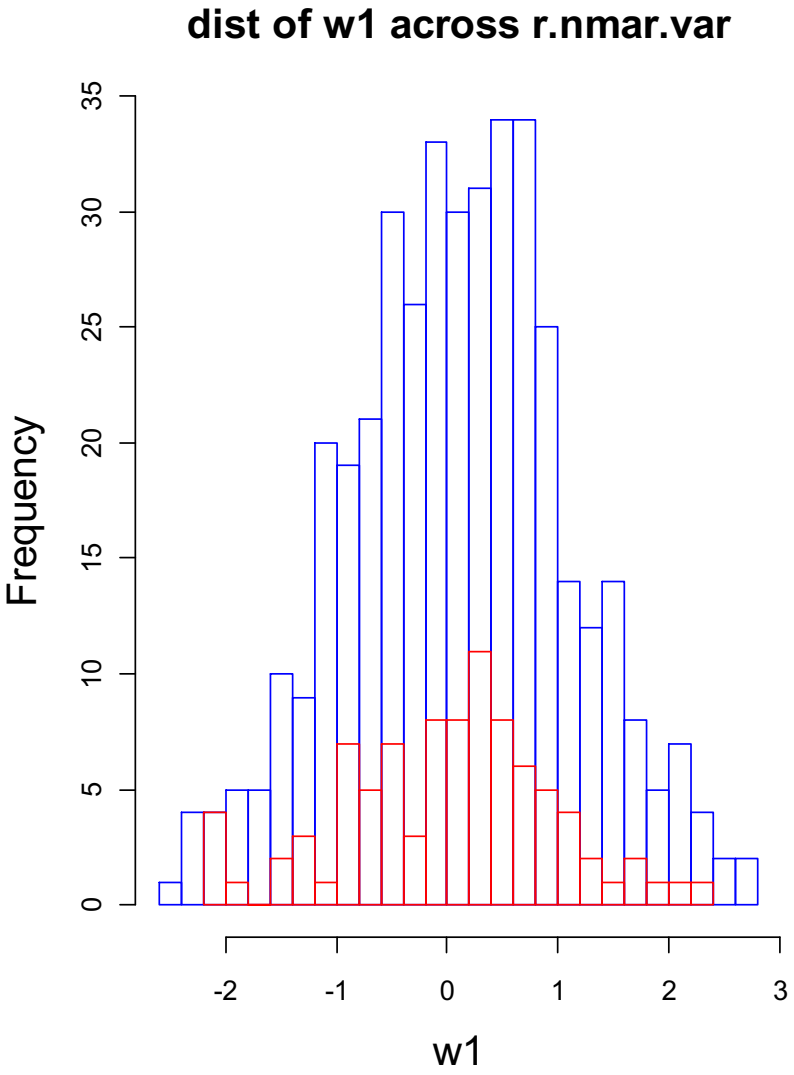


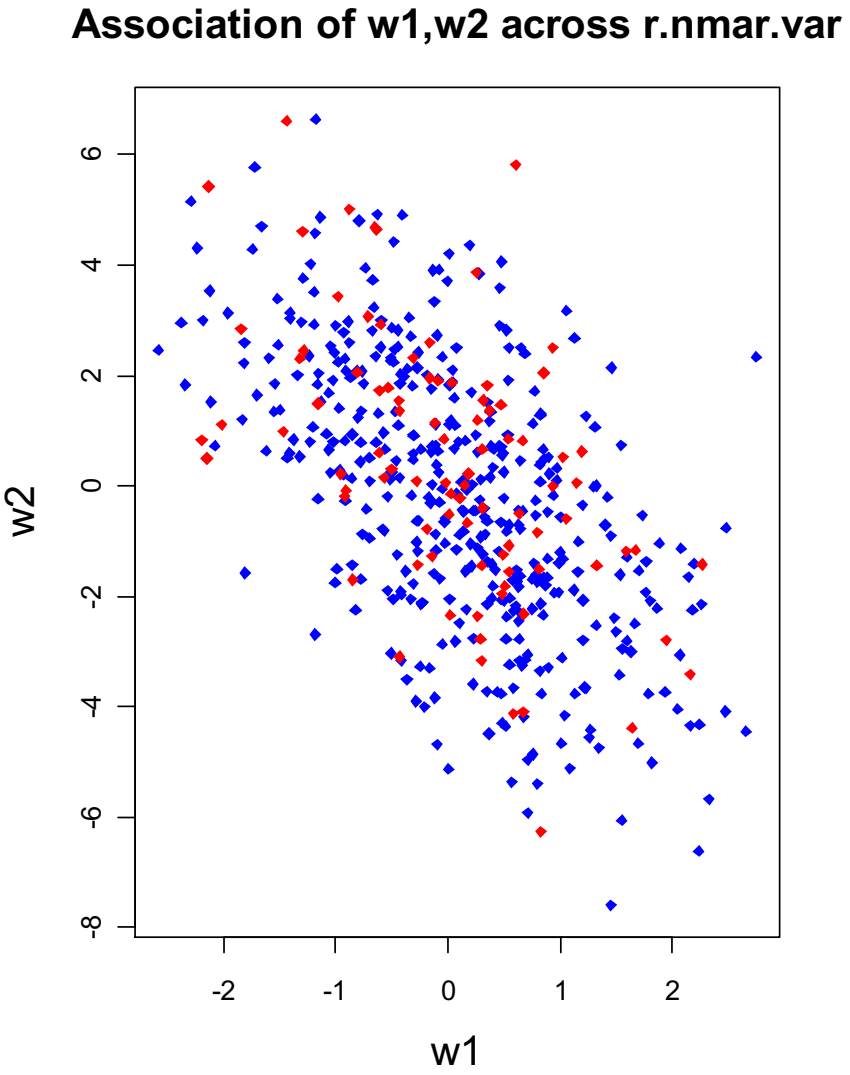
Checks regarding missingness of Y_{mar}





Checks regarding missingness of Y_{nmar}





Information regarding missing data patterns (our simulated data)

- We don't find violations of MCAR in Y_{mcar} when looking at its missingness relative to W_1 and W_2 (just as we would expect)
- We do find violations of MCAR in Y_{mar} when looking at its missingness relative to W_1 and W_2. However, when we condition on W_2 (by regressing both W_1 and Y_{mar} on W_2 and using the residuals) we remove the dependence between R (missing indicator for Y_{mar}) and W_1.
- What is particularly worrisome in the extreme case of NMAR (where Y_{nmar} is uncorrelated with all the other variables) is that the data will not obviously even violate MCAR!

What can we learn?

- We may be able to rule out MCAR
- Also, there is Little's test where H_0: MCAR, so small p-value rejects the null, but no known results for Type II errors.
- We can never say for sure that something is MCAR though
- We cannot distinguish empirically between MAR and MNAR

Example on p. 32 of Enders

- data contain the responses from 400 college-aged women on 10 questions from the Eating Attitudes Test (EAT)
- The data set also contains an anxiety scale score, a variable that measures beliefs about Western standards of beauty (e.g., high scores indicate that respondents internalize a thin ideal of beauty), and body mass index (BMI) values.
- Author intentionally created MCAR, MAR and NMAR missing values in different variables
- We test for MCAR using Little's test from the package BaylorEdPsych
- Not surprisingly, the null hypothesis is rejected!

What can we do when we have missing data?

Methods that throw away data
Methods that impute data

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

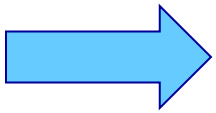
- Complete cases
- Complete variables
- Pairwise deletion (for special purposes)

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

- Imputation

Complete Cases

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



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)
- 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

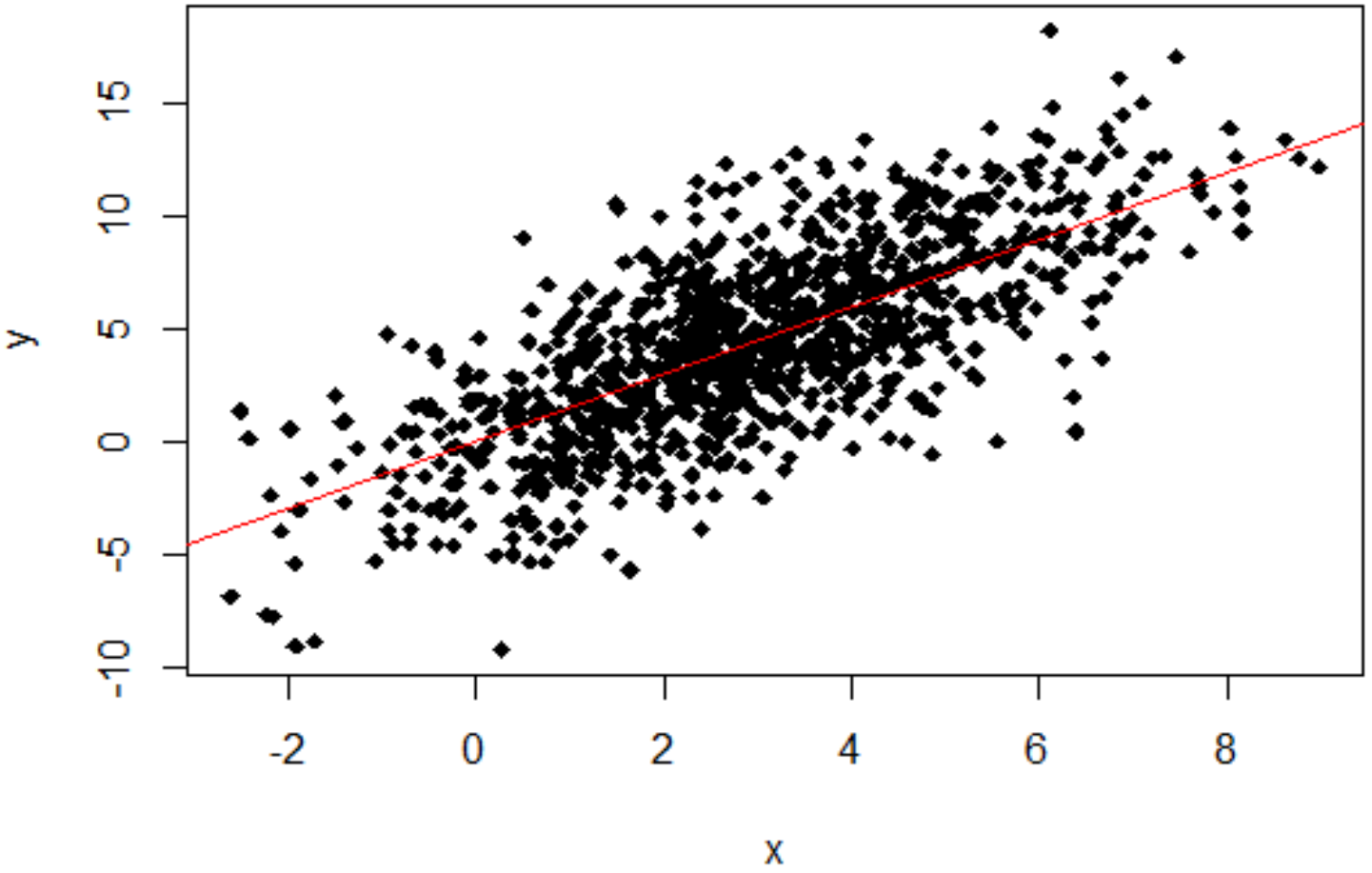
Complete case assumptions

- We generally talk about complete case analyses rest on the MCAR assumption
- Two ways of thinking about
  - Intuition 1: complete case sample is a random subsample of the complete sample
  - Intuition 2: each person in the sample has an equal probability of having been dropped from the sample
- Some issues apply (discussed on next slide)

Special cases where CCA is unbiased

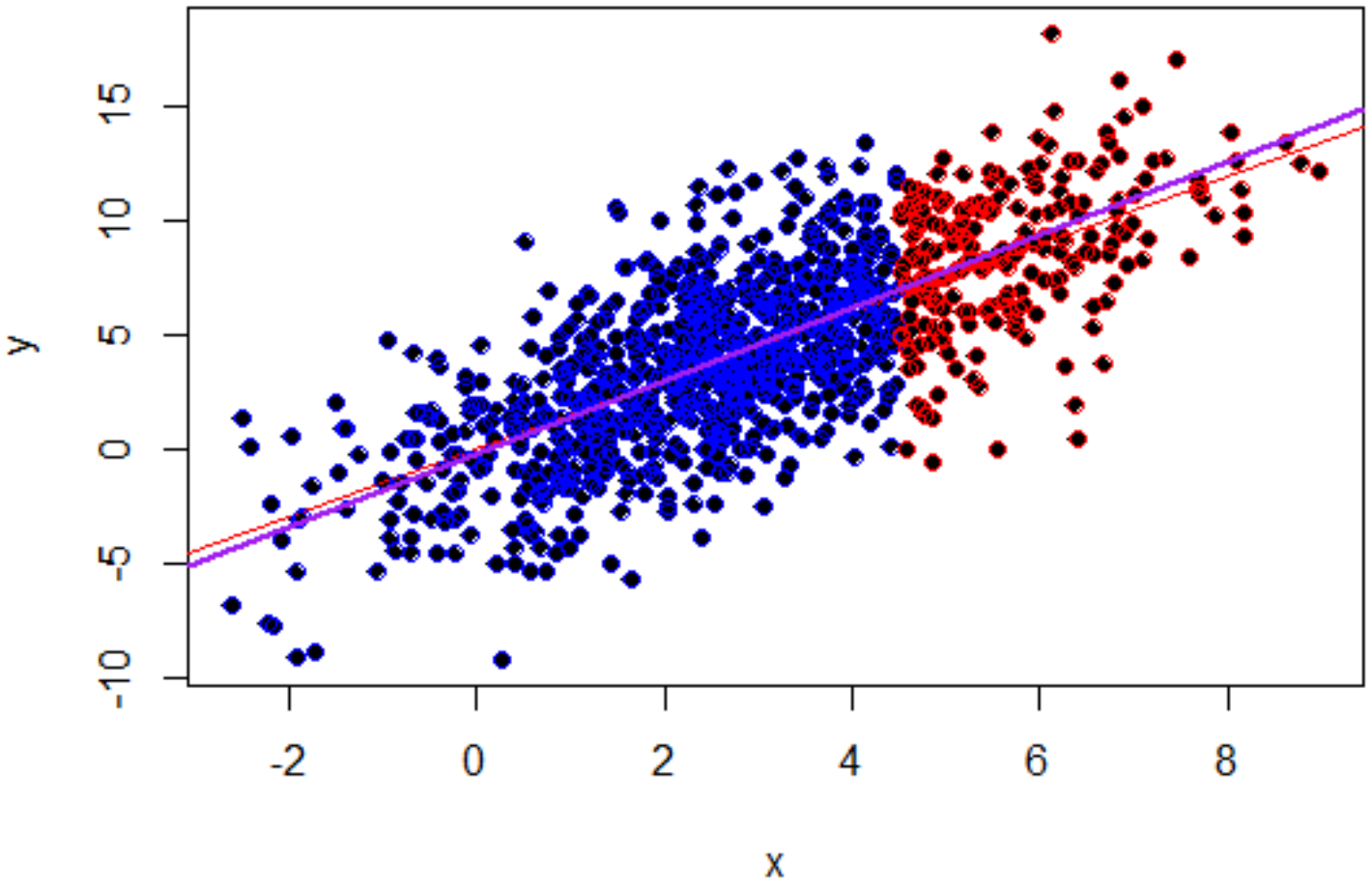
- Suppose the goal is to regress Y_1 on Y_2, … ,Y_k with a linear regression. If the missingness does not depend on Y_1 then complete cases will do fine *even if the MDM is NMAR* (e.g. if $\Pr(R_{Y_k}) = g(Y_k)$).
- Suppose the goal is to regress Y_1 on Y_2 … Y_k: with a *logistic* regression. If *either* binary Y_1 or binary Y_j is missing and that missingness depends only on Y_1 then a CCA will be unbiased
- These estimates could still be quite inefficient (see for instance King et al. 2001)
- However, when there are complicated missing data patterns, the sample size can get reduced very quickly to a point when it will be difficult to use the remaining data to fit an appropriate model or to extrapolate to the rest of the original sample

Illustration of CCA and regression



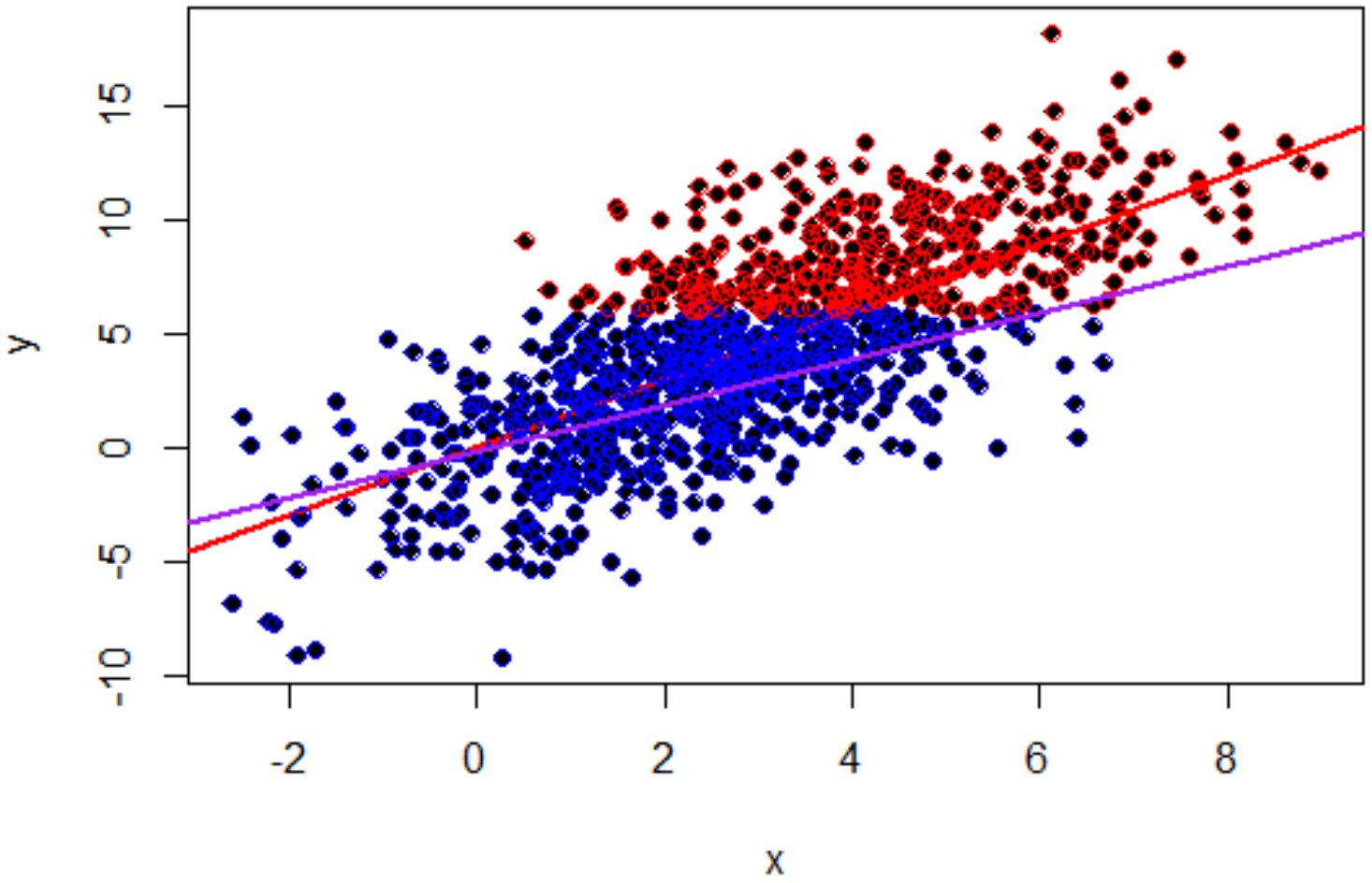
CCA and regression: just fitting to observed (blue) points

Missingness due to x



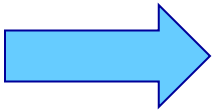
CCA and regression: just fitting to observed (blue) points

Missingness due to y



Complete variables

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



X_1	X_2
0	2
0	1
0	2
0	3
1	5
1	4
1	6
1	6

Complete variables

- This throws out all *variables* that have any missing data
- So full sample size is retained, but we may be left with only a few variables which aren't of sufficient substantive interest
- Makes it difficult to meet assumptions need for other types of analyses

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!

HW 1 (first step towards final project)

- Find suitable dataset with missing data (around 20% missing);
- Should contain *both* numerical and categorical variables;
- There should be at least 3-4 variables;
- There should be at least 50 observations (cases);
- At least one *numerical* variable should be *completely observed*;
- Decide what analysis you want to do with the data;
- If having difficulty finding data with missing values, use complete data and we will learn how to create missing.