Generating missing values with ampute

Rianne Schouten

1. Department of Methodology and Statistics, University Utrecht
2. DPA Professionals Data Science Excellence Program

May 18, 2017

Overview

- What is amputation?
- Characteristics of missing data problems
- Stepwise Univariate Amputation
- Multivariate Amputation

```
require(mice)
?ampute
```

What is amputation?

Amputation is the generation of missing values in complete data

Evaluation of a missing data methodology:

- Simulate complete data set
- Generate missing values
- Deal with missing data with new method
- Compare statistical inferences between the original, complete data set and after dealing with the missing values

Characteristics of missing data problems

MCAR : Missingness is not related to X or Y at all

MAR : Missingness is related to X but not to Y

MNAR : Missingness is related to Y

MCAR : $Pr(Y = missing) = 0.5$

MAR : $logit(Pr(Y = missing)) = X$

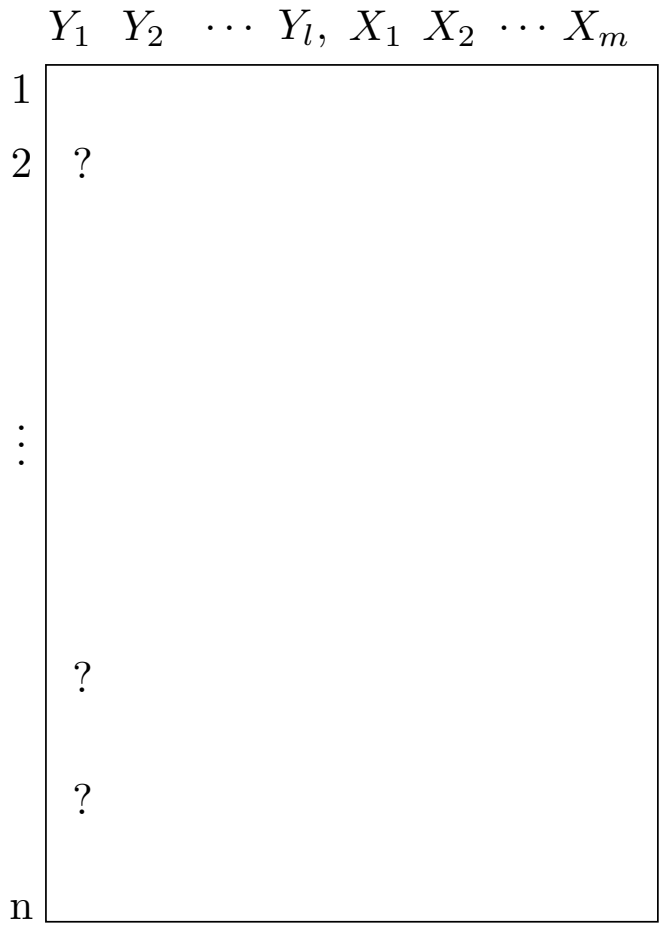
MNAR : $logit(Pr(Y = missing)) = Y$

Characteristics of missing data problems

- Mechanism
- Proportion
- Severity
- Which variables
- Influence of observed data
- Combinations

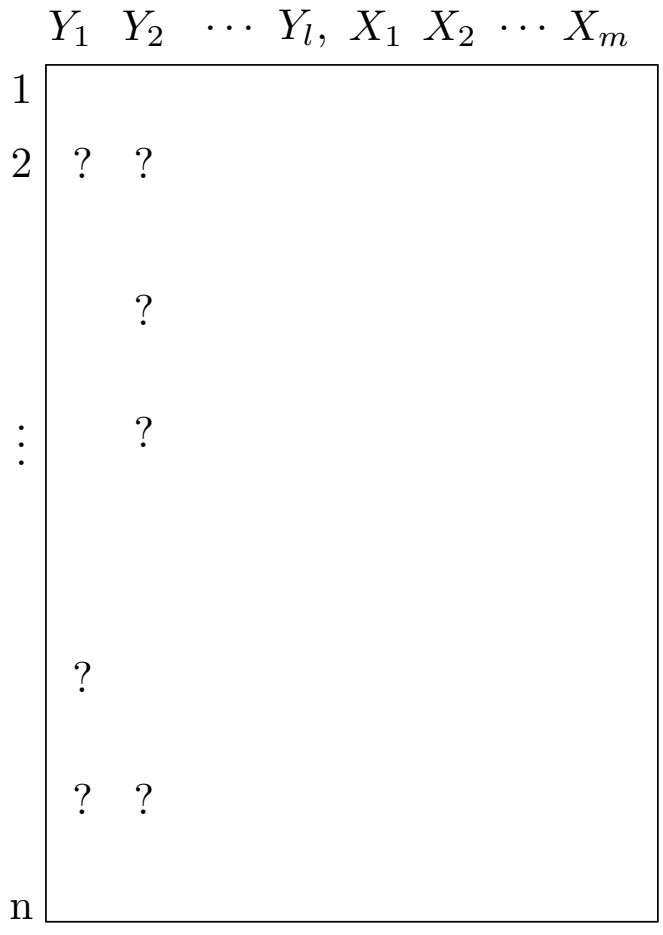
Stepwise Univariate Amputation

1. Y_1



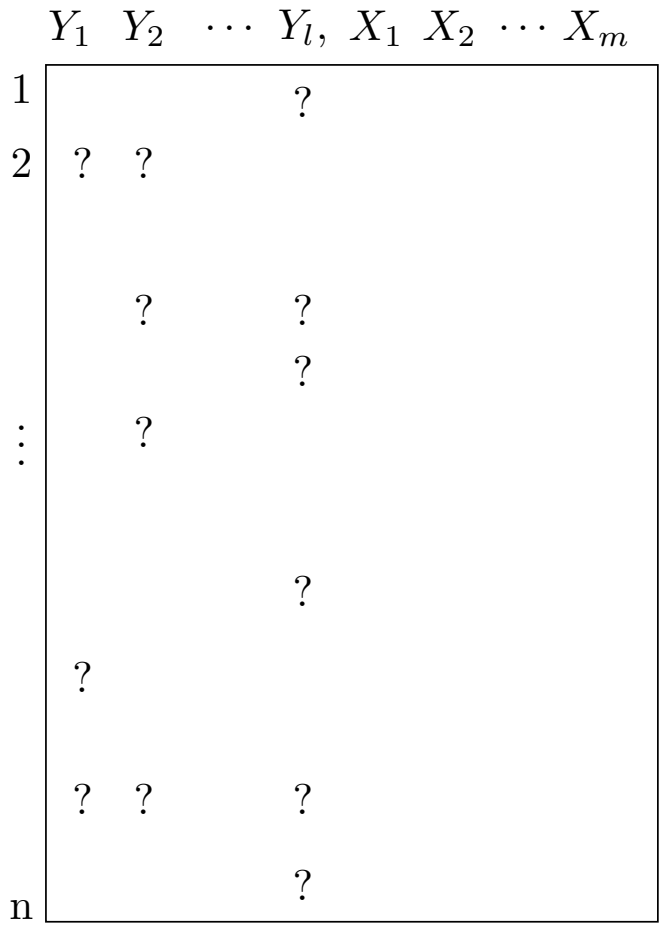
Stepwise Univariate Amputation

1. Y_1
2. Y_2

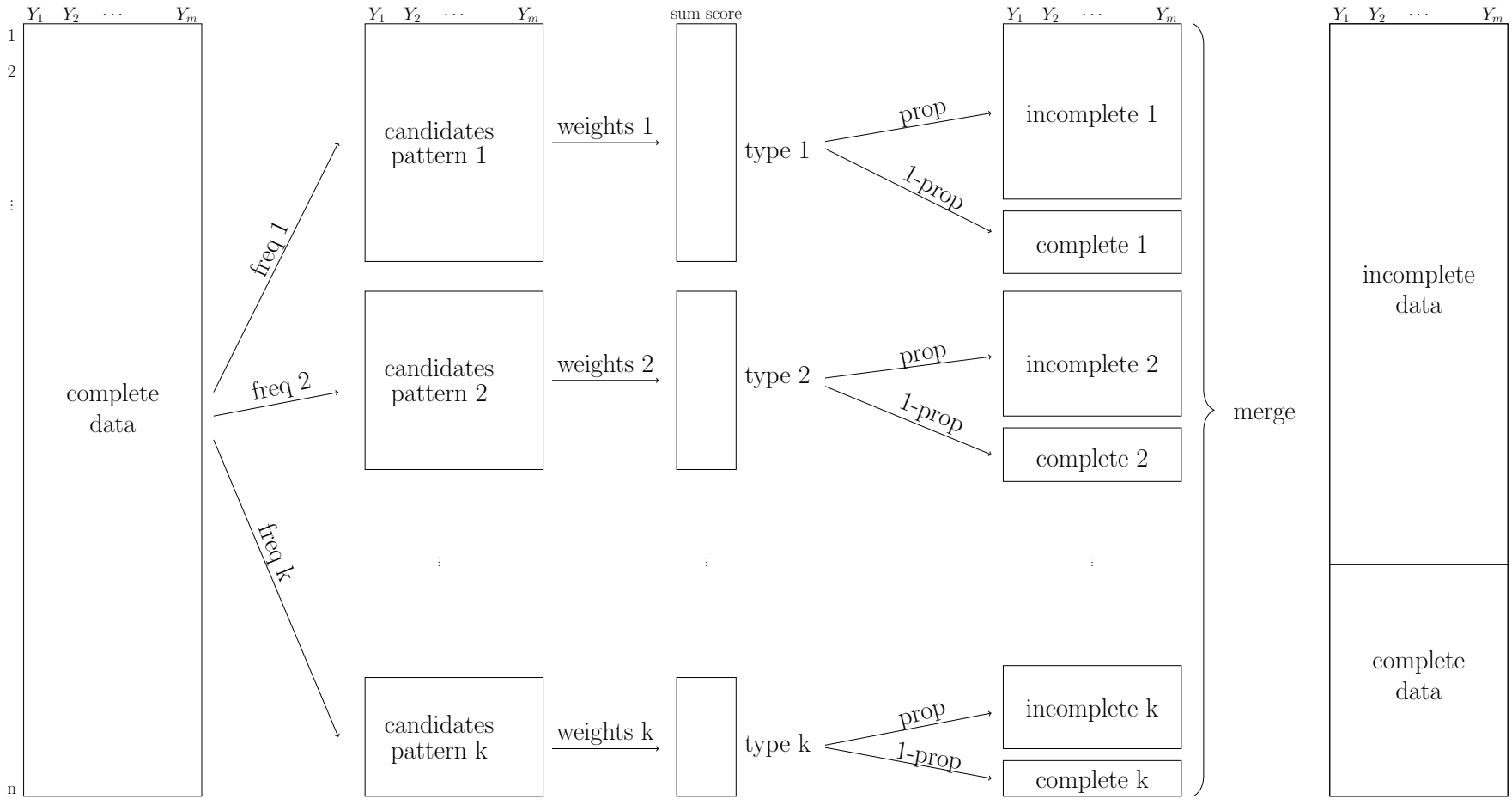


Stepwise Univariate Amputation

1. Y_1
2. Y_2
3. ...
4. Y_l



Multivariate Amputation



Multivariate Amputation: Missing data patterns

A missing data pattern is a specific combination of variables with missing values and variables without missing values.

0: incomplete variable

1: complete variable

Multivariate Amputation: Missing data patterns

```
data <- mvrnorm(mu = c(5, 5, 10),
                Sigma = matrix(c(1, 0.5, 0.5,
                                 0.5, 1, 0.5,
                                 0.5, 0.5, 1), 3, 3),
                n = 1000)
data <- as.data.frame(data)
names(data) <- c("y1", "y2", "x")
head(data)
```

```
##         y1       y2        x
## 1 4.587067 5.364096 10.333693
## 2 3.070874 5.139578  9.588529
## 3 4.557591 5.771920  9.834172
## 4 4.709714 3.481258  9.890152
## 5 5.080492 6.236113 10.939586
## 6 4.255989 4.980705  9.431149
```

Multivariate Amputation: Missing data patterns

```
mypat <- matrix(c(0, 0, 1,
                  0, 1, 1),
                2, 3, byrow = TRUE)
amp <- ampute(data, patterns = mypat)
inc_data <- amp$amp
```

Multivariate Amputation: Missing data patterns

```
inc_data[1:12, ]
```

```
##           y1       y2         x
## 1         NA       NA 10.333693
## 2         NA 5.139578  9.588529
## 3         NA       NA  9.834172
## 4         NA       NA  9.890152
## 5         NA 6.236113 10.939586
## 6   4.255989 4.980705  9.431149
## 7         NA 5.711130  9.607115
## 8         NA       NA 11.388759
## 9         NA       NA 10.708886
## 10 5.046014 5.876691 10.369283
## 11 4.143609 4.770560  9.499410
## 12 5.193741 4.729886  8.890983
```

Multivariate Amputation: Proportion and Frequency

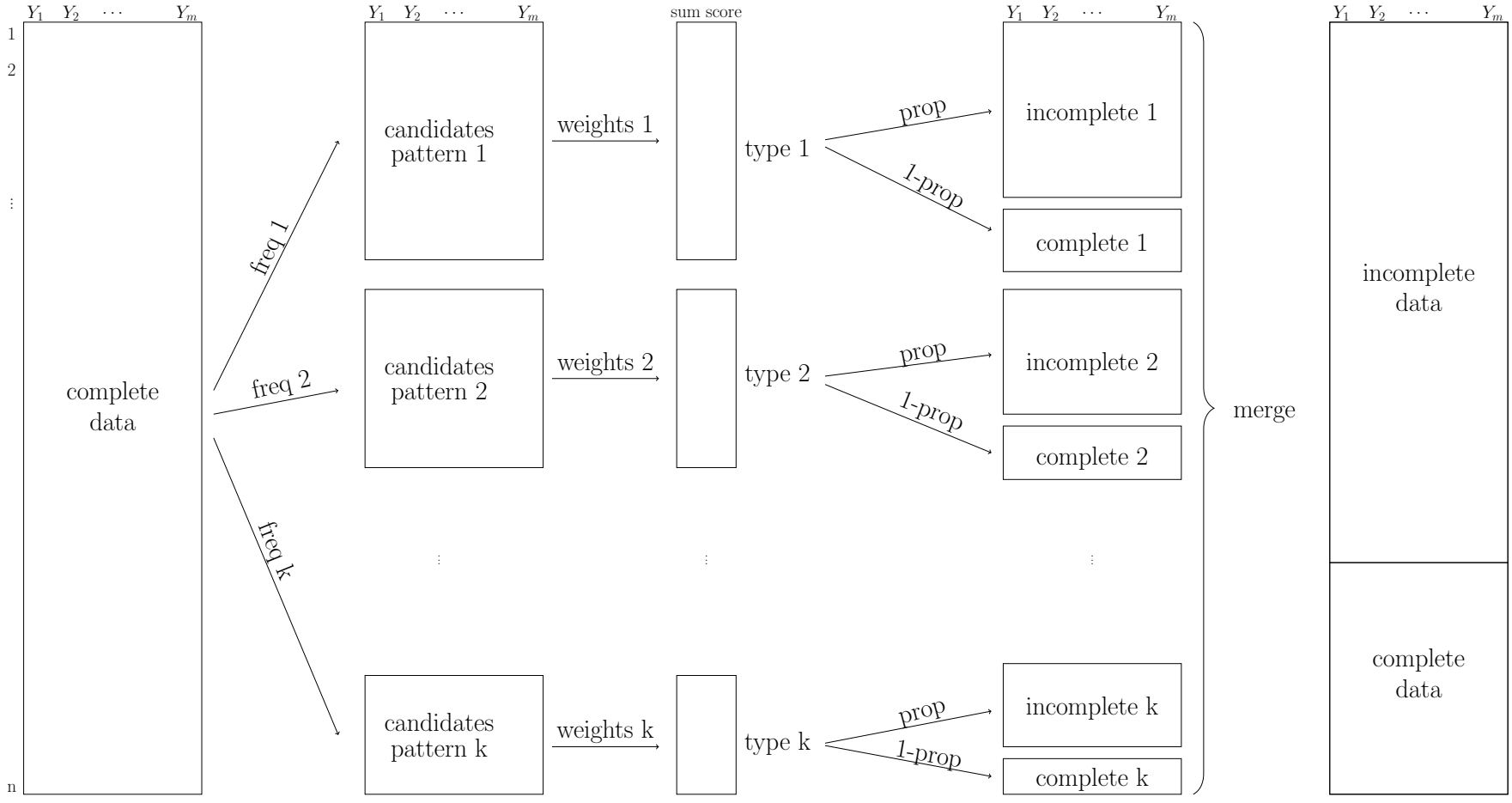
```
amp$prop
```

```
## [1] 0.5
```

```
amp$freq
```

```
## [1] 0.5 0.5
```

Multivariate Amputation: Proportion and Frequency



Multivariate Amputation: Proportion and Frequency

```
myprop <- 0.8
myfreq <- c(0.8, 0.2)
amp <- ampute(data, patterns = mypat,
              prop = myprop, freq = myfreq)
md.pattern(amp$amp)
```

```
##     x  y2  y1
## 197 1   1   1    0
## 156 1   1   0    1
## 647 1   0   0    2
##     0 647 803 1450
```

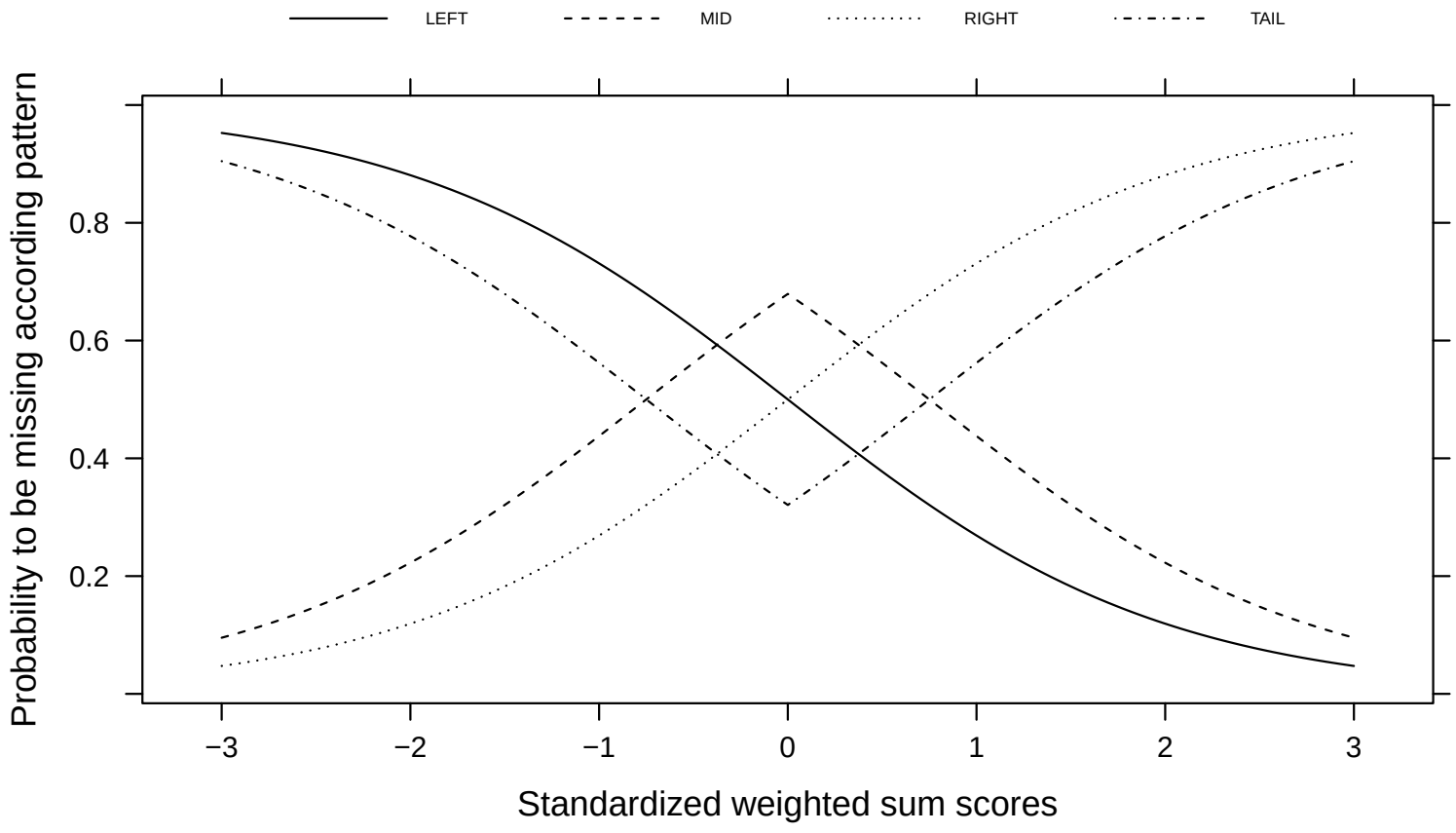
Multivariate Amputation

```
ampute(data, prop = 0.5, patterns = NULL, freq =
NULL, mech = "MAR", weights = NULL, cont = TRUE, type
= NULL, odds = NULL, bycases = TRUE, run = TRUE)
```

```
amp$patterns
```

```
##   y1 y2 x
## 1  0  0 1
## 2  0  1 1
```

Multivariate Amputation: Distribution functions




```
amp$type
```

```
## [1] "RIGHT" "RIGHT"
```

Multivariate Amputation: Weighted sum scores

- Missing values in multiple variables

$$\begin{array}{c} \\ P_1 \\ P_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y_1 & Y_2 & X \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array}$$

- Based on multiple variables

  $wss_i = w_{1,1} \cdot y_{1i} + w_{1,2} \cdot y_{2i} + w_{1,3} \cdot x_i$ if case i is in pattern 1

$$\begin{array}{c} \\ W_1 \\ W_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y_1 & Y_2 & X \end{array} \\ \begin{bmatrix} 1 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

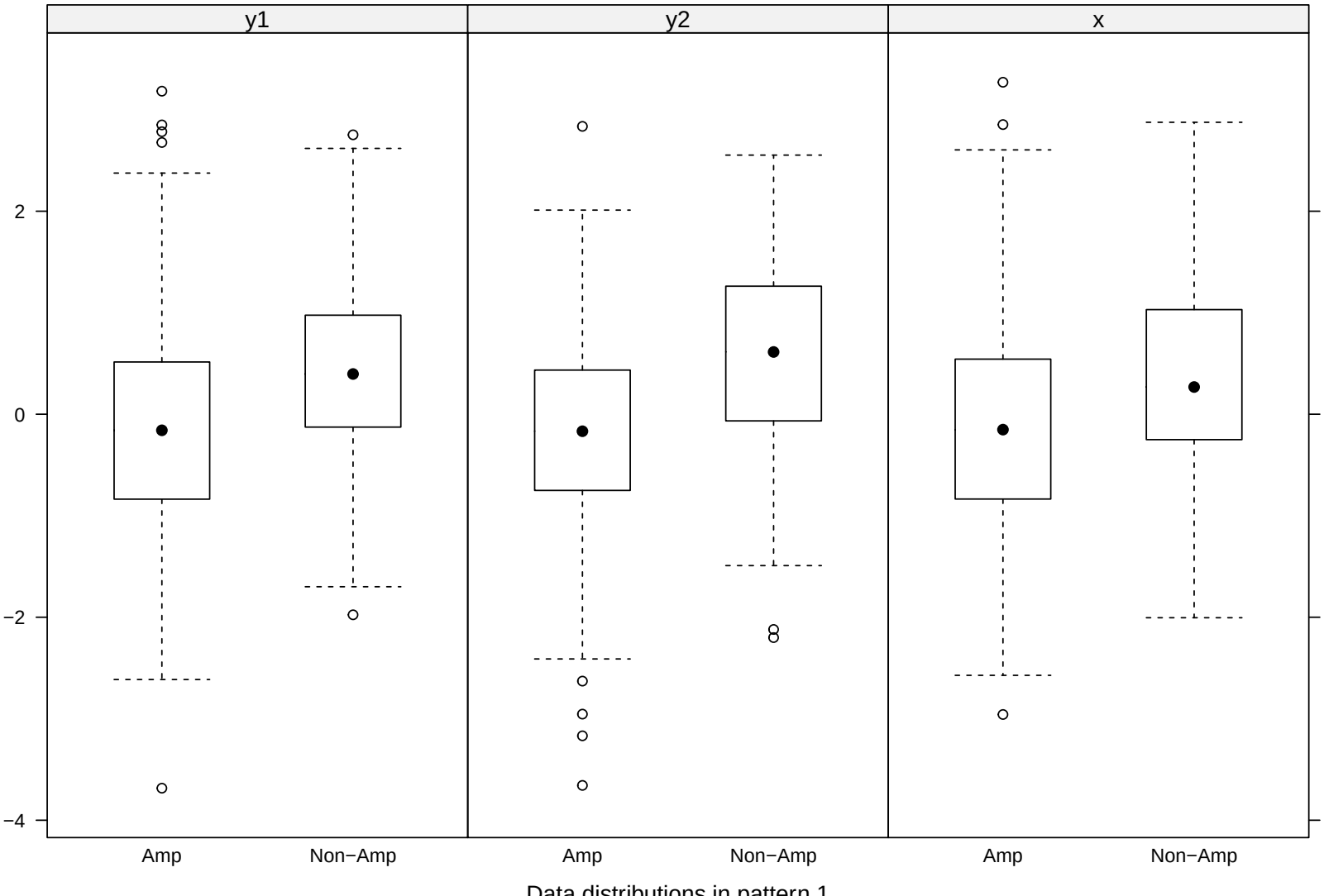
Multivariate Amputation: Weighted sum scores

```
myweights <- matrix(c(1, 5, 0,
                      0, 0, 1),
                    2, 3, byrow = TRUE)

amp <- ampute(data, patterns =  mypat,
              prop = myprop, freq = myfreq,
              weights = myweights,
              type = c("LEFT", "MID"))
```

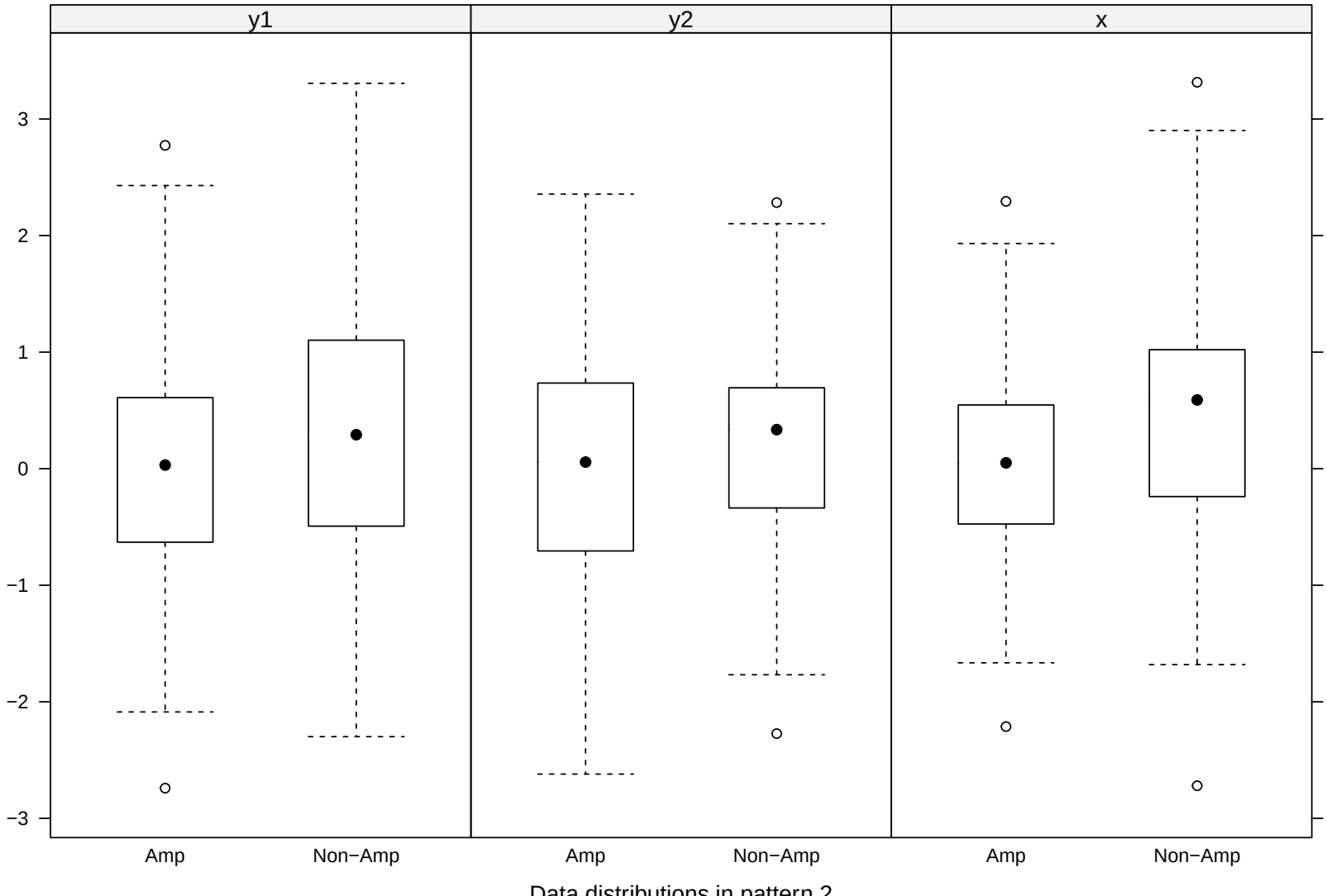
Multivariate Amputation

```
## $`Boxplot pattern 1`
```



Multivariate Amputation

```
## $`Boxplot pattern 2`
```



Multivariate Amputation

- Mechanism
- Proportion
- Severity
- Which variables
- Influence of observed data
- Combinations

```
ampute(data, prop = 0.5, patterns = NULL, freq =
NULL, mech = "MAR", weights = NULL, cont = TRUE, type
= NULL, odds = NULL, bycases = TRUE, run = TRUE)
```

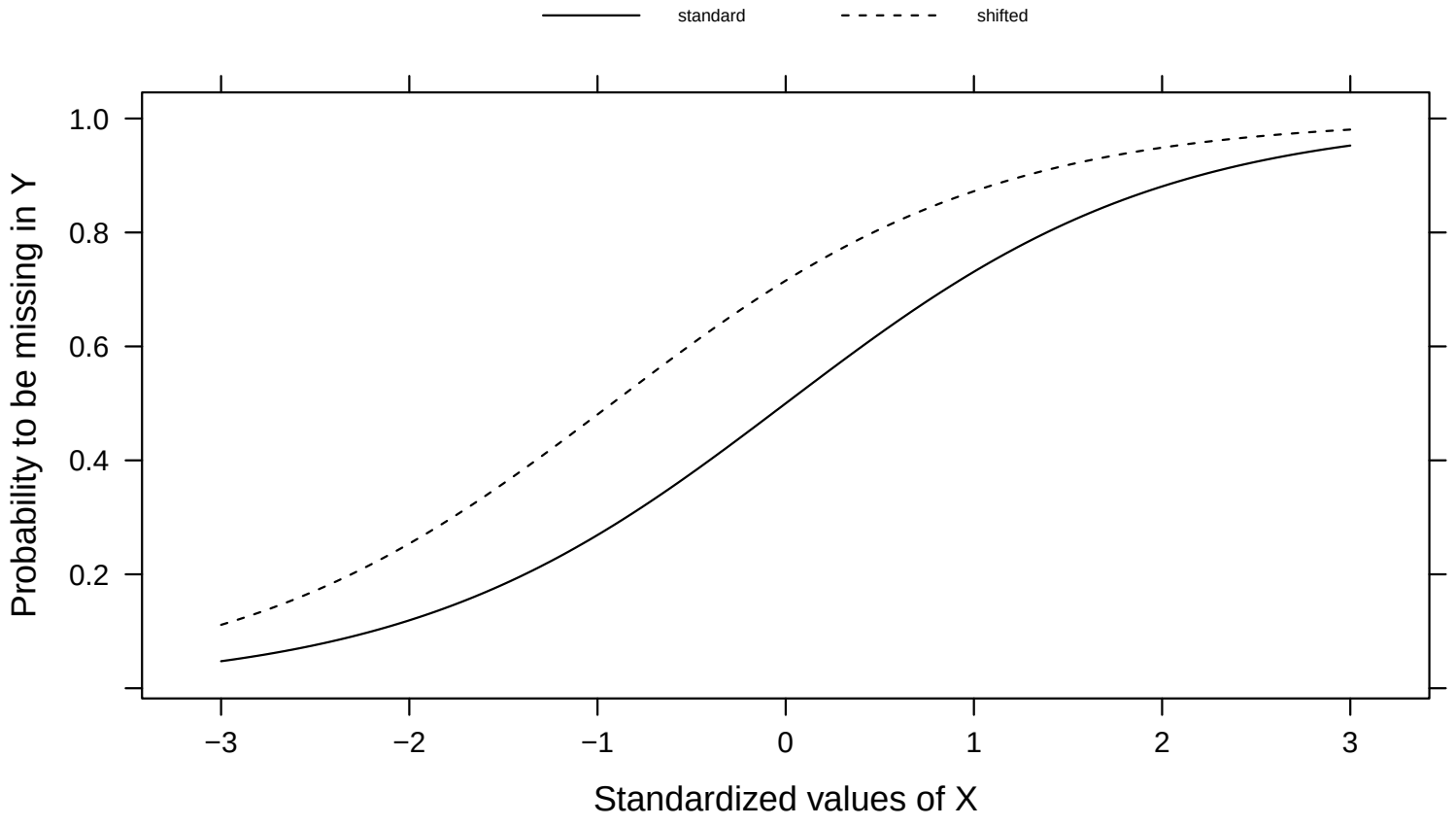
Documentation

Rianne Schouten, Peter Lugtig, Jaap Brand, Gerko Vink (2017) Generate missing values with `ampute`, available from: `https://github.com/RianneSchouten/Amputation_with_Ampute/tree/master/Manuscript%20article`

Rianne Schouten, Peter Lugtig, Gerko Vink (2017) Generating missing values for simulation purposes: A multivariate amputation procedure. Under review. Available from: `https://github.com/RianneSchouten/mice/tree/master/vignettes`

```
require(mice)
?ampute
```

Additional slides

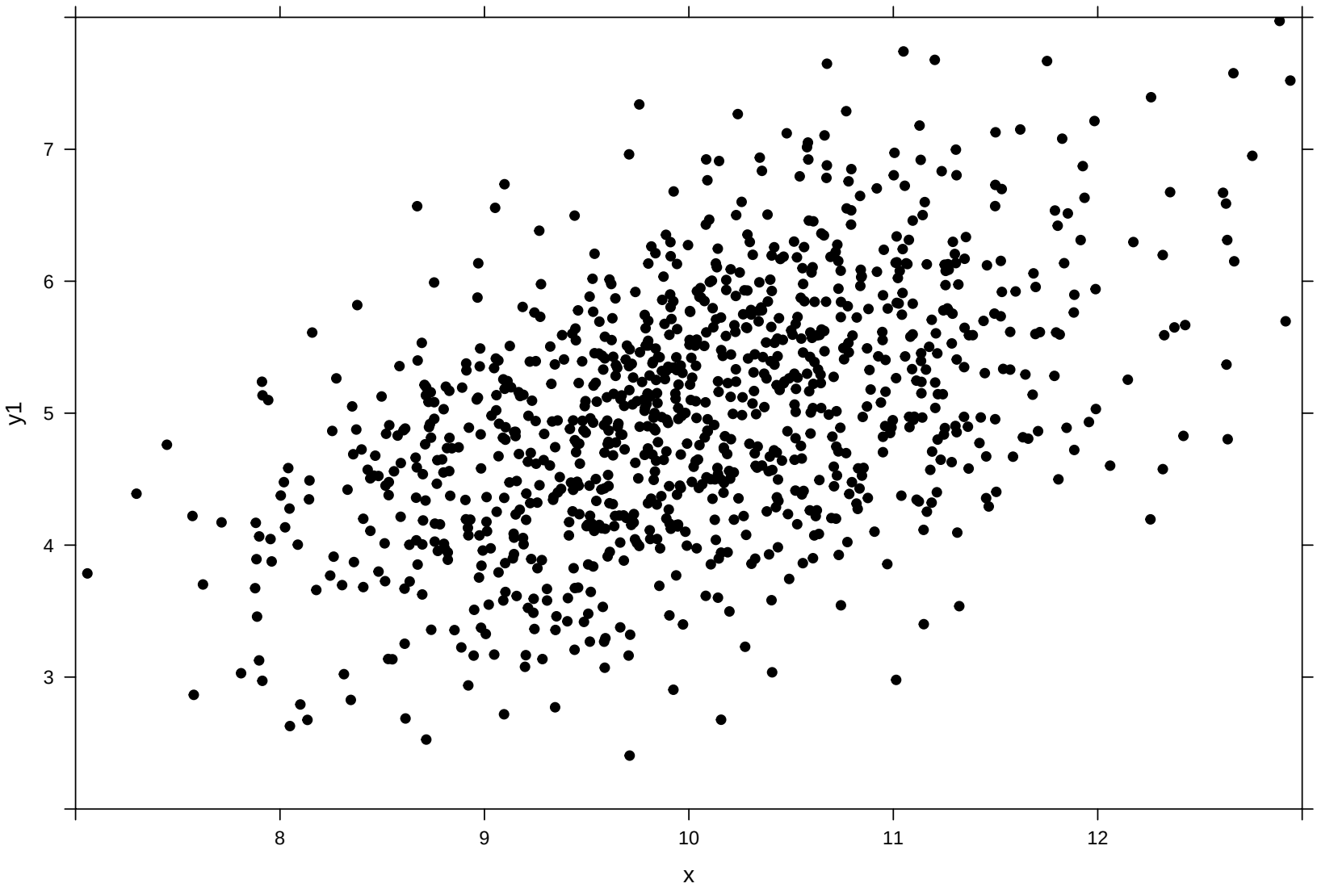


Additional slides

Table 1: Generation of MAR missingness on 2 variables with standard and shifted stepwise univariate amputation (SUA) and multivariate amputation (MA)

		%mis		complete case analysis			multiple imputation		
cor	condition	int	obt	bias	ciw	cov	bias	ciw	cov
	standard SUA	50	29	-0.146	0.144	0.028	-0.002	0.156	0.940
0.5	shifted SUA	50	50	-0.233	0.172	0.000	-0.007	0.204	0.917
	MA with `ampute`	50	50	-0.207	0.172	0.002	-0.005	0.193	0.936

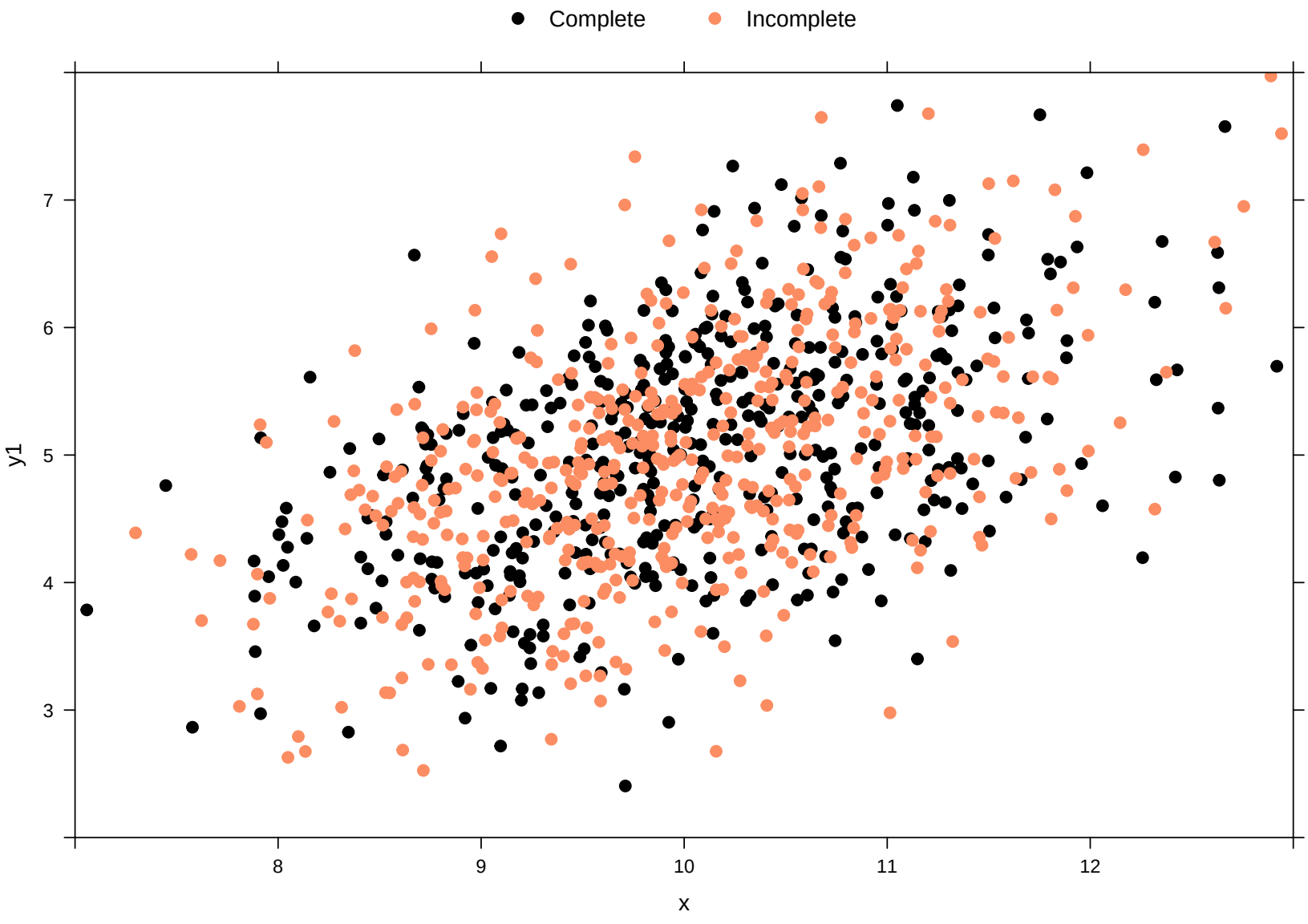
Characteristics of missing data problems



Characteristics of missing data problems: MCAR

```
mcar_data <- ampute(data, pat = c(0, 1, 1),
                    mech = "MCAR")$amp
```

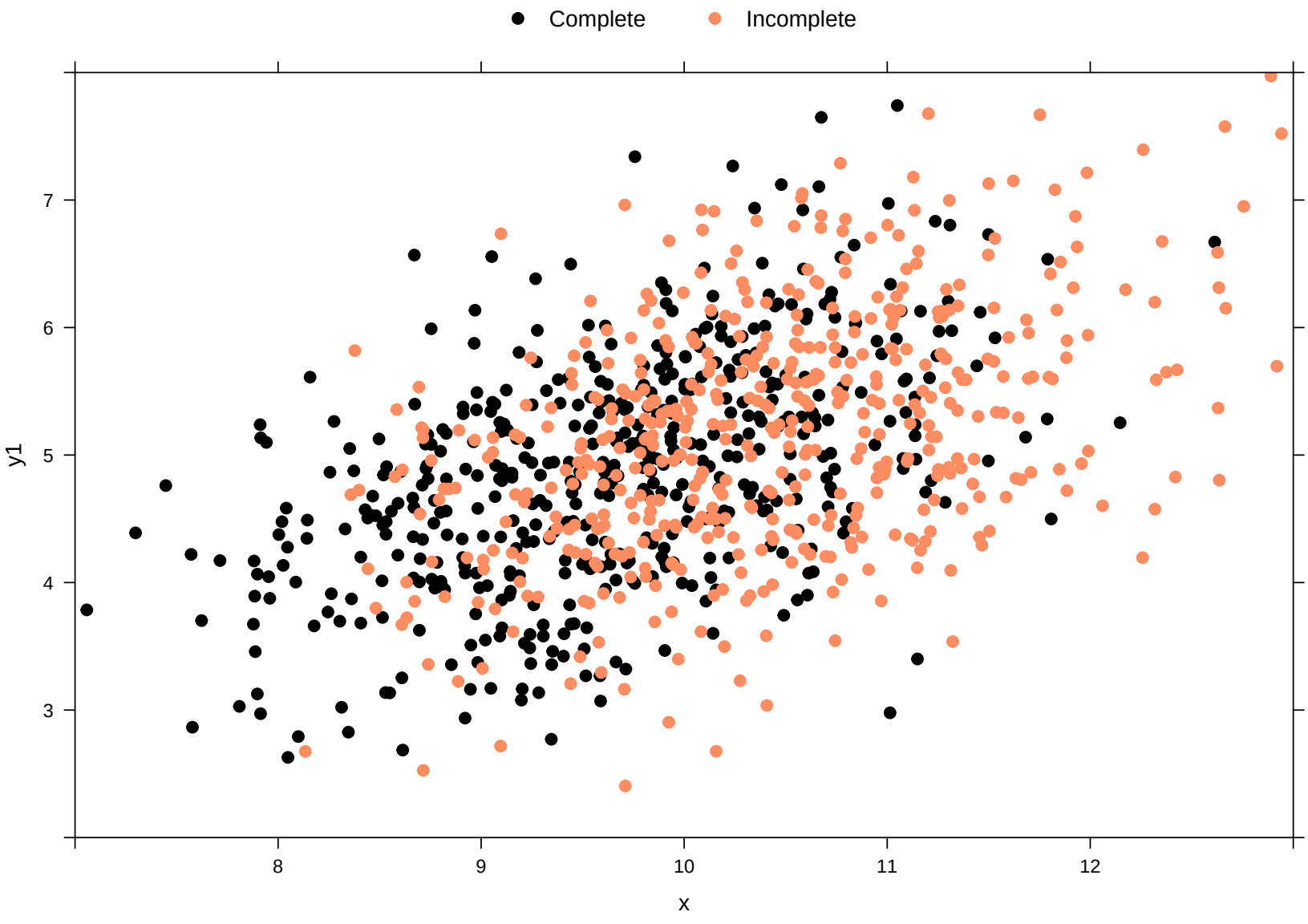
Characteristics of missing data problems: MCAR



Characteristics of missing data problems: MAR

```
mar_data <- ampute(data, pat = c(0, 1, 1),
                   weights = c(0, 0, 1),
                   mech = "MAR")$amp
```

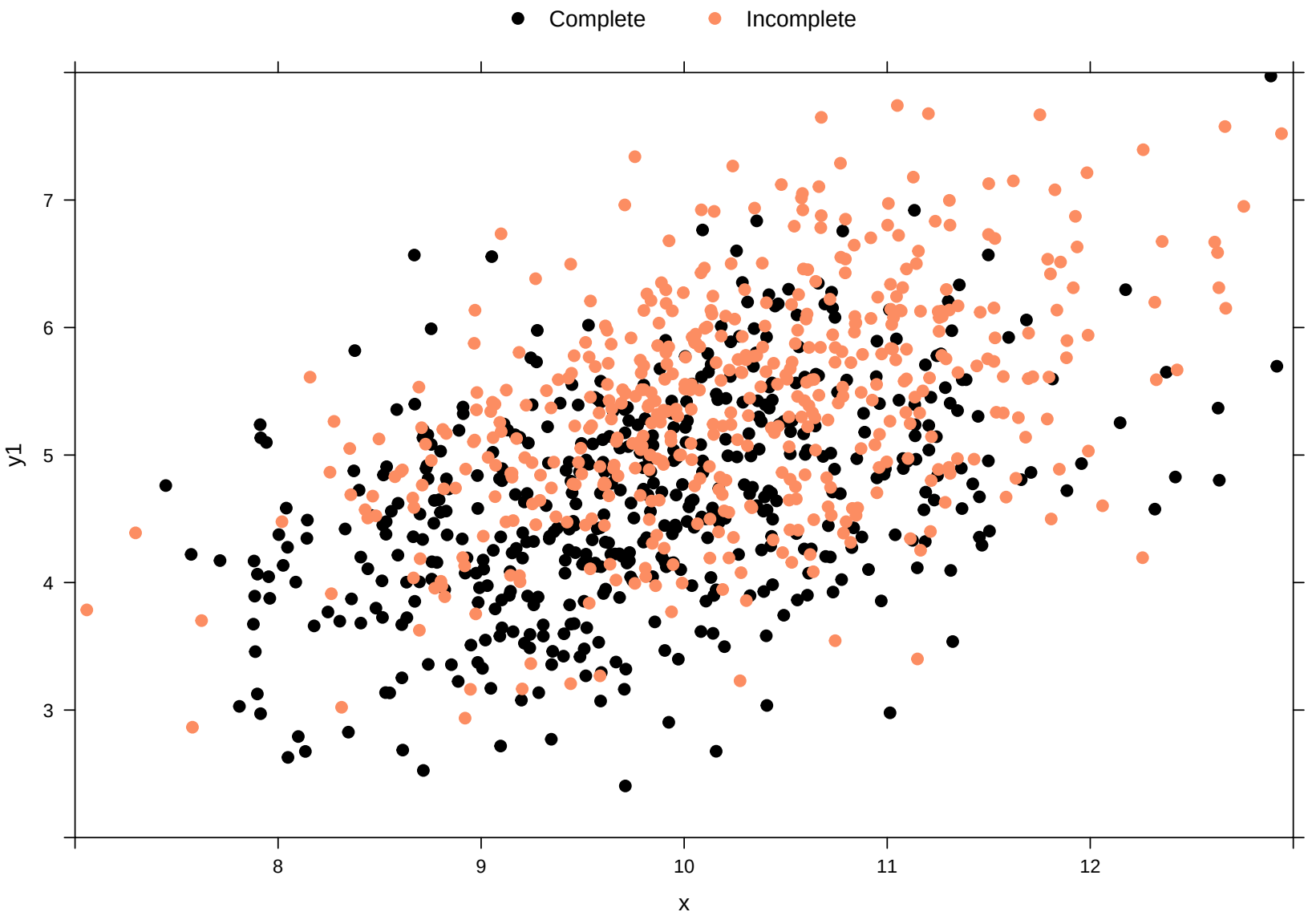
Characteristics of missing data problems: MAR



Characteristics of missing data problems: MNAR

```
mnar_data <- ampute(data, pat = c(0, 1, 1),
                    weights = c(1, 0, 0),
                    mech = "MNAR")$amp
```

Characteristics of missing data problems: MNAR



Documentation

Rianne Schouten, Peter Lugtig, Jaap Brand, Gerko Vink (2017) Generate missing values with `ampute`, available from: `https://github.com/RianneSchouten/Amputation_with_Ampute/tree/master/Manuscript%20article`

Rianne Schouten, Peter Lugtig, Gerko Vink (2017) Generating missing values for simulation purposes: A multivariate amputation procedure. Under review. Available from: `https://github.com/RianneSchouten/mice/tree/master/vignettes`

```
require(mice)
?ampute
```