Generating missing values for simulation purposes: A multivariate amputation approach

Rianne Schouten

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

July 7, 2017

Amputation

Amputation is the generation of missing values in complete data

Overview:

- Why?
- What?
- How?

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

Amputation: Why?

Evaluation of missing data methodologies:

1. Simulate complete data set
2. Generate missing values
3. Deal with missing data with new method
4. Compare statistical inferences

But also:

- Planned missing data survey designs
- Investigate measurement errors
- See effect missing data on your analyses

Amputation: What?

- Proportion
- Amputed variables

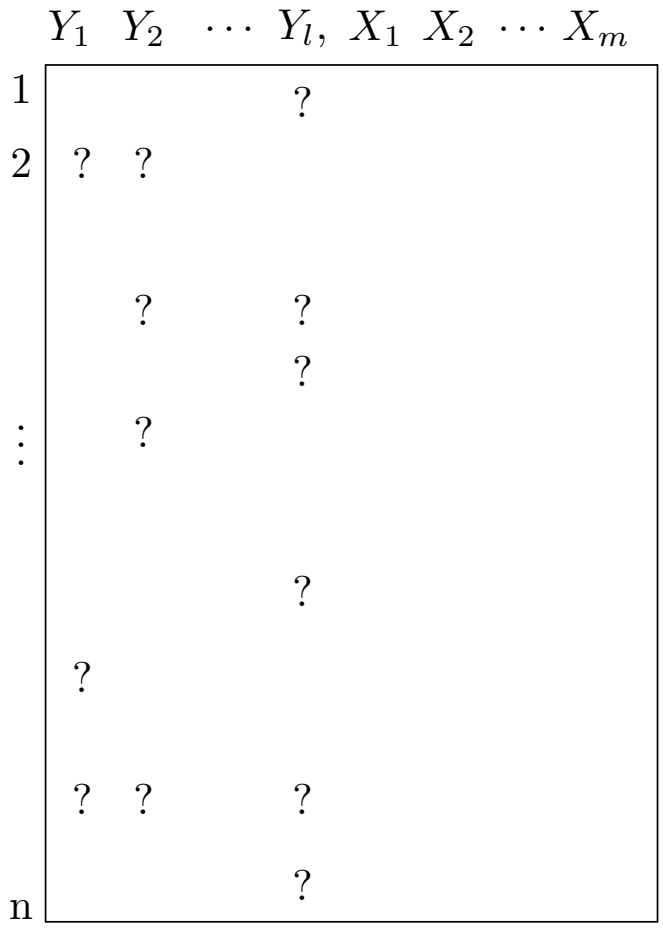
Amputation: What?

- Proportion
- Amputed variables
- Mechanism

MCAR : Mis not related to X or Y at all

MAR : Mis related to X but not to Y

MNAR : Mis related to Y

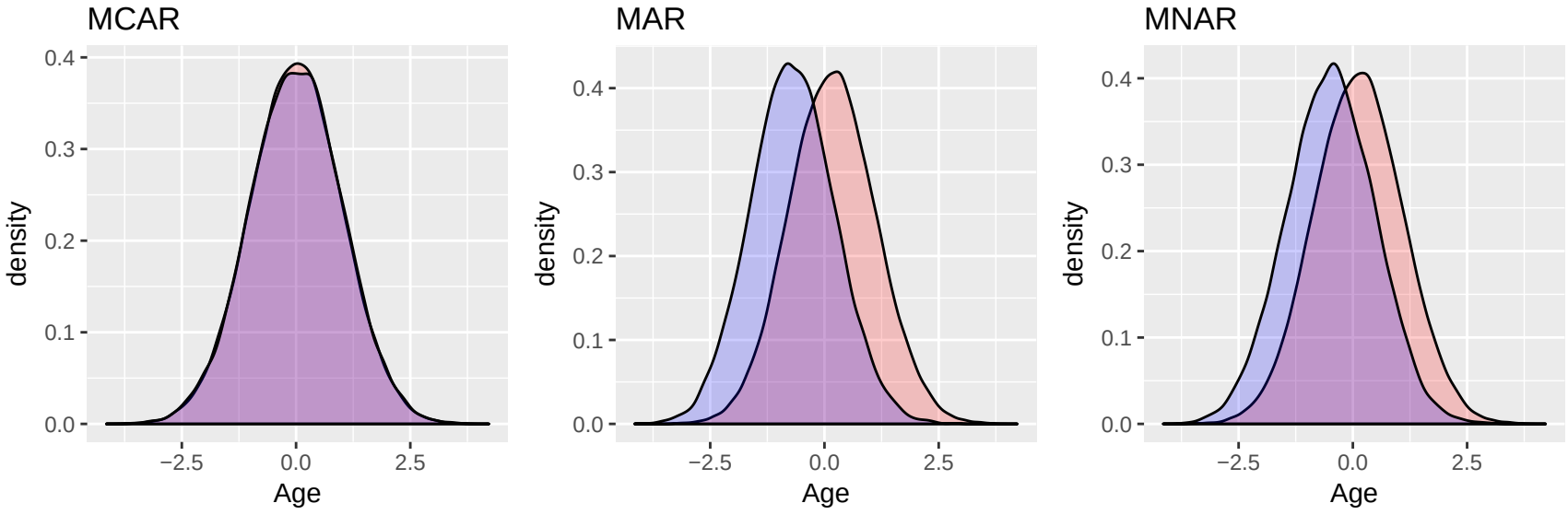


Amputation: What?

```
# Customers Phone Company
head(customer_data)
```

```
##       Income    Minutes         Age
## 1         NA  1.9237723  0.4174930
## 2 -0.6322071 -0.2409715  0.2492411
## 3 -0.9443980 -1.2539681 -0.5233141
## 4         NA  3.1223591  1.2705289
## 5 -0.4703926  0.5594291 -0.9440021
## 6 -0.3342031 -0.7998914 -0.1937294
```

Amputation: What?



$$\text{MCAR} : Pr(Income = missing) = 0.5$$

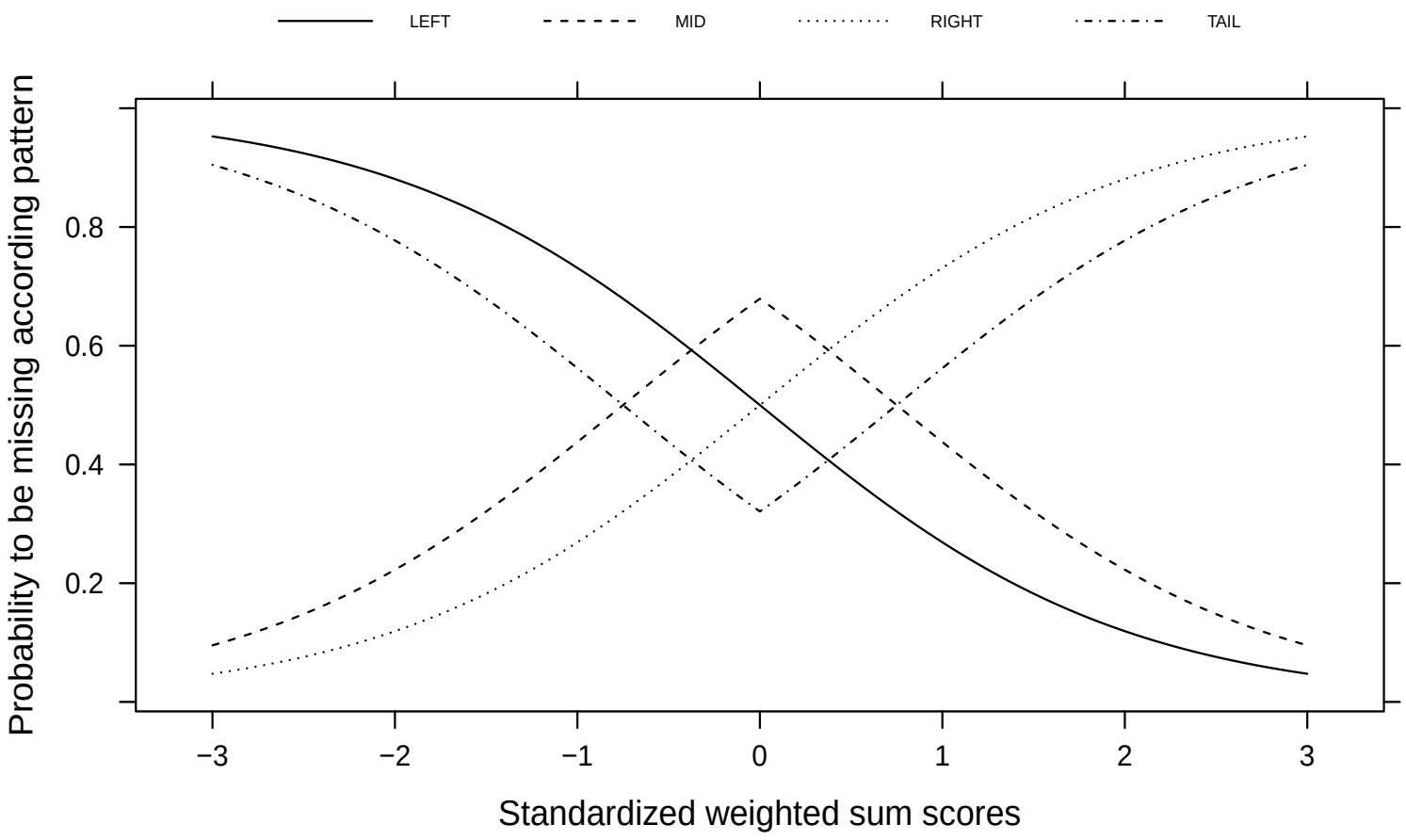
$$\text{MAR} : Pr(Income = missing) = Age$$

$$\text{MNAR} : Pr(Income = missing) = Income$$

Amputation: What?

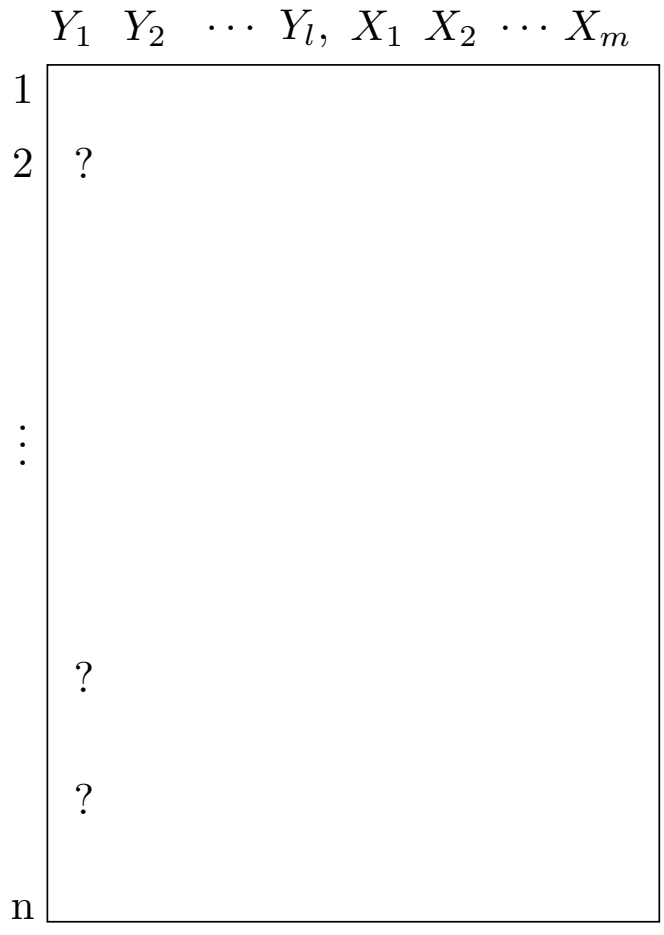
- Proportion
- Amputed variables
- Mechanism
- Influencing variables
- Severity
- Missingness distribution

Amputation: What?



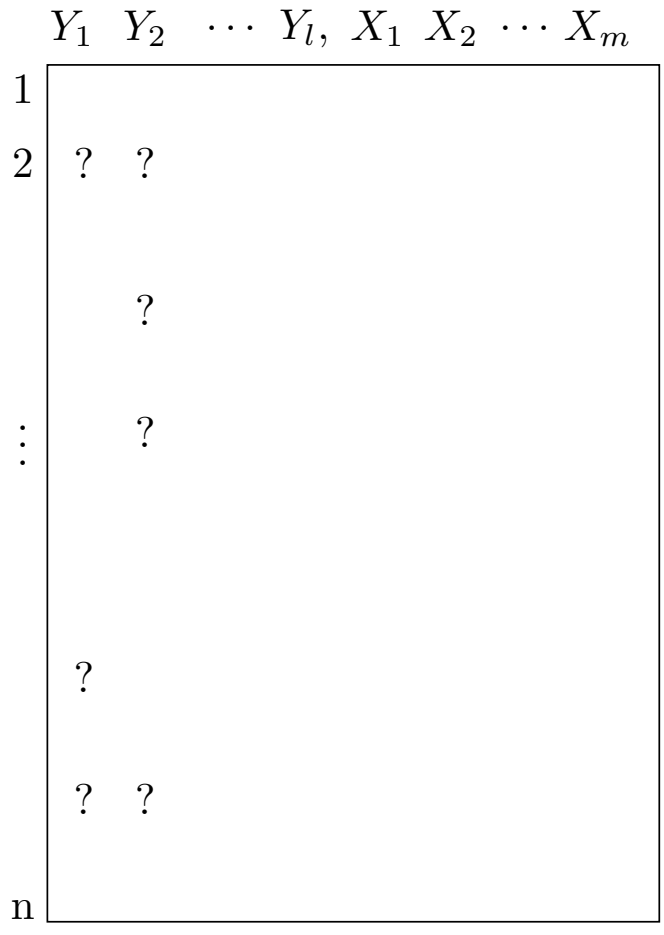
Amputation: How?

1. Y_1



Amputation: How?

1. Y_1
2. Y_2



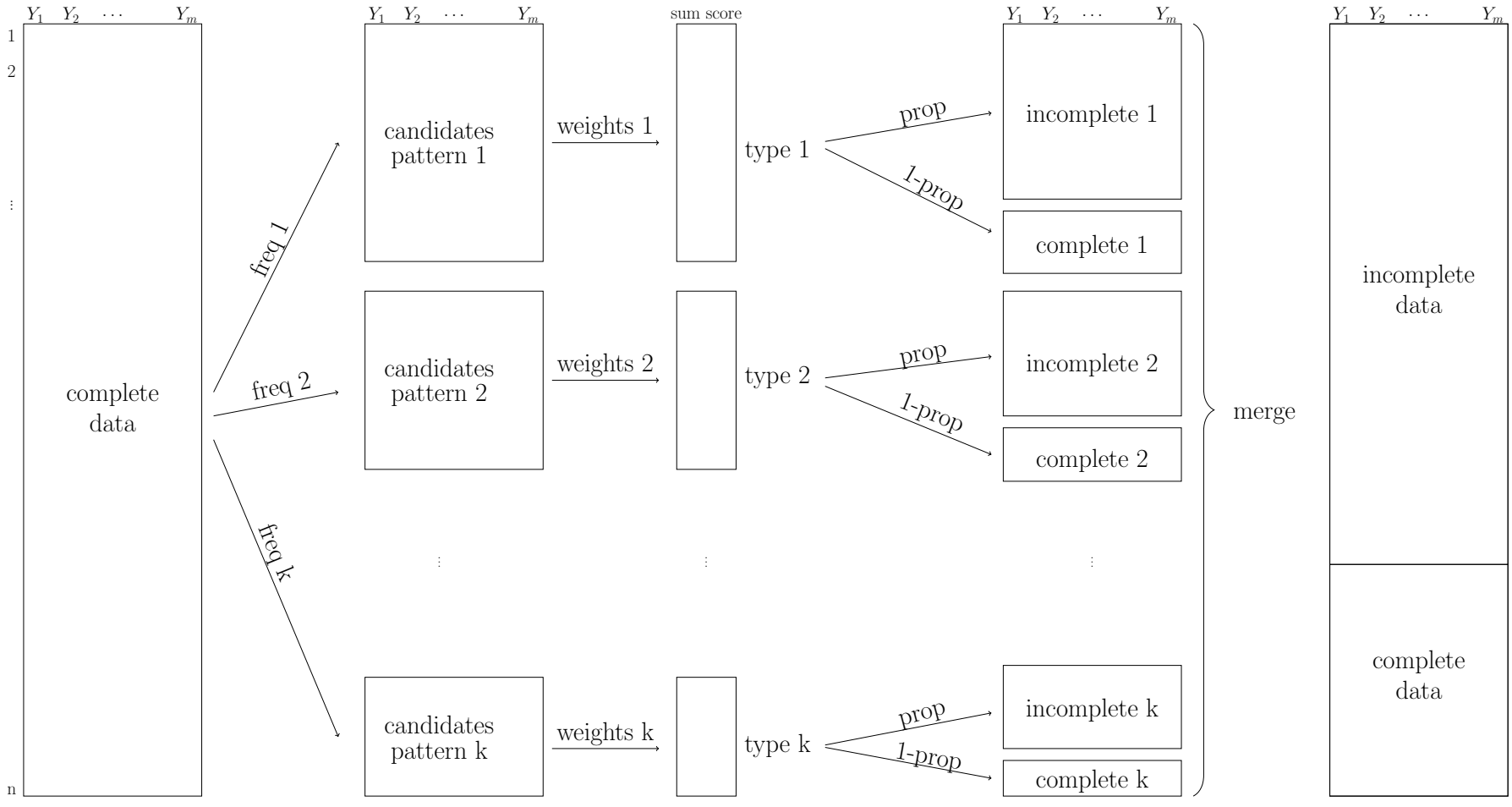
Amputation: How?

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

	Y_1	Y_2	$\cdots$	Y_l,	X_1	X_2	$\cdots$	X_m
1				?				
2	?	?						
		?		?				
				?				
$\vdots$		?						
				?				
	?							
	?	?		?				
n				?				

Amputation with ampute

Multivariate Amputation:



Amputation with ampute

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

```
amp <- ampute(data)
class(amp)
```

```
## [1] "mads"
```

Amputation with ampute

```
head(amp$amp)
```

```
##       Income    Minutes         Age
## 1  0.8250012  1.9237723          NA
## 2 -0.6322071 -0.2409715  0.2492411
## 3 -0.9443980 -1.2539681 -0.5233141
## 4         NA  3.1223591  1.2705289
## 5 -0.4703926  0.5594291 -0.9440021
## 6 -0.3342031         NA -0.1937294
```

Documentation

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

Vignette:

https://github.com/RianneSchouten/Amputation_with_Ampute

Contact:

Rianne Schouten, r.m.schouten@uu.nl, rianne.schouten@dpa.nl

Universiteit Utrecht

DPA | IT

Additional slides

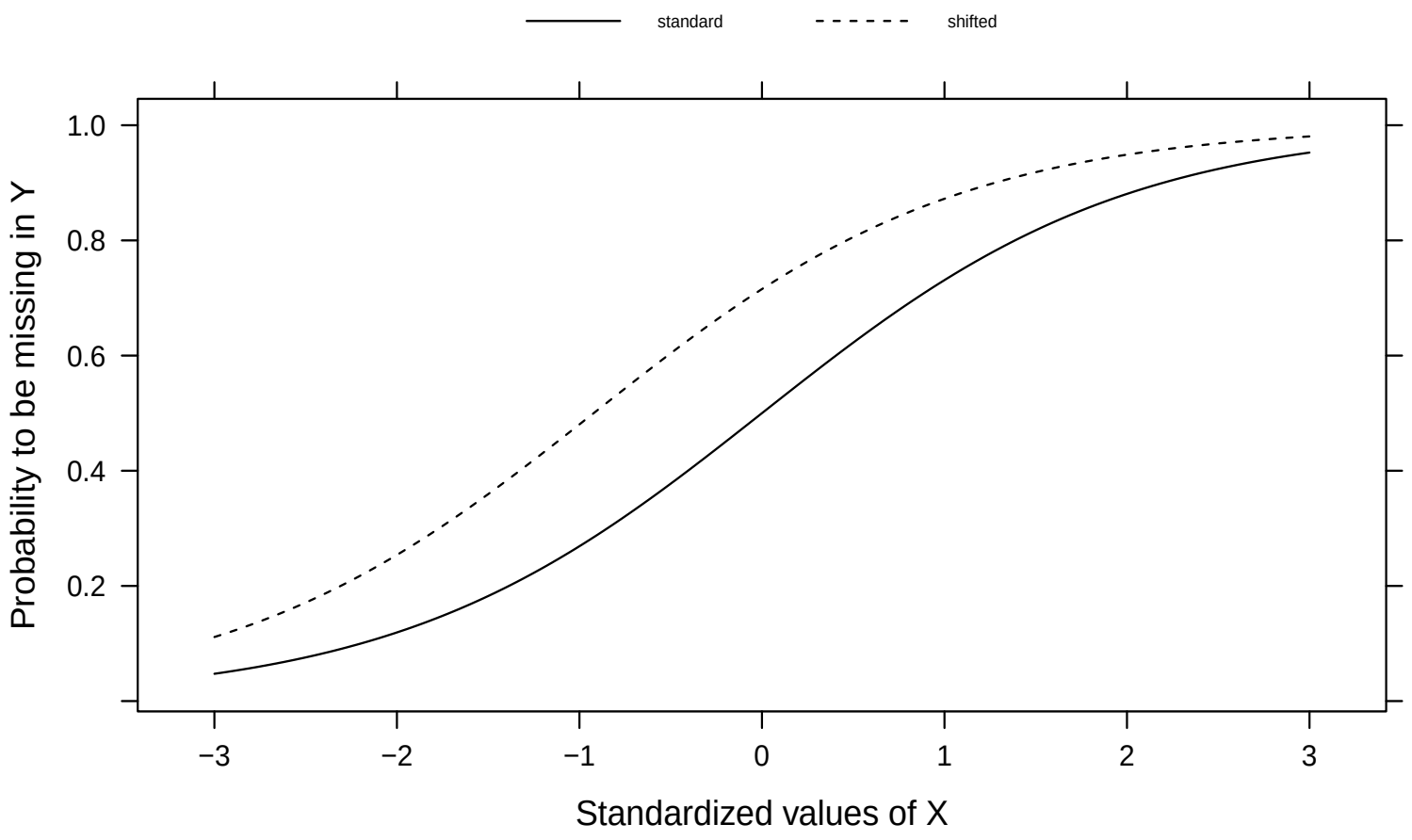
A missing data pattern is a specific combination of variables with missing values and variables without missing values.

0: incomplete variable

1: complete variable

```
mypat <- matrix(c(0, 0, 1,
                  0, 1, 1),
                nrow = 2, byrow = TRUE)
```

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
0.5	standard SUA	50	29	-0.146	0.144	0.028	-0.002	0.156	0.940
	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