GSO international workshop

Mathematatic, biostatistics and epidemiology of cancer

Modeling and simulation of clinical trials

HANDLING MISSING DATA

Gregory GUERNEC[1], Valerie GARES[1,2]

[1]*UMR1027 INSERM*
UNIVERSITY OF TOULOUSE III

[2]*NHMRC Clinical Trials Centre*
UNIVERSITY OF SYDNEY

INTRODUCTION AND DEFINITIONS

Simulated complete dataset

- 150 subjects and 9 variables

Variables	*Types*	*Details*	
ID	*Identification*	*From 1 to 150*	
method	*Factor in 2 groups [A/B]*	$p_A = 0.5$, $p_B = 0.5$	
educ	*Education levels in 3 groups [1 / 2 / 3]*	$p_1 = 0.3$, $p_2 = 0.5$, $p_3 = 0.2$	
sex	*Gender in 2 groups [F/M]*	$p_F = 0.6$, $p_M = 0.4$	
x	*A continuous covariate*	*Normal, M = 212, SD = 10*	
y_0	*Dependent continuous variable [time 0]*	*Normal, M = 130, SD = 10*	
y_1	*Dependent variable time 1*	*Normal, M = 156.5, SD = 13.3*	*From method, sex, x,* y_0
y_2	*Dependent variable time 2*	*Normal, M = 195.5, SD = 16.4*	*From method, sex, x,* y_0
y_3	*Dependent variable time 3*	*Normal, M = 221.7, SD = 18.8*	*From method, sex, x,* y_0

- Create missing data
  - Only in y_0, y_1, y_2, y_3
  - Rest is completely observed

Simulated complete dataset

- Visualisation of the first 20 IDs …

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	134.6131	158.7102	201.8993	229.4077
2	B	3	F	205.4844	128.5105	152.9126	191.3709	219.0251
3	B	3	M	220.8016	130.4923	156.0304	195.2262	221.5212
4	B	2	M	208.1906	147.7882	176.9052	221.7698	252.6078
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	160.9798	202.3034	230.9175
7	A	2	M	215.7308	136.5162	165.6273	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.7470	212.1675
9	A	2	F	221.9858	152.7514	182.1968	228.9192	260.2738
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.2210	225.7538
12	A	1	M	209.8337	149.9582	180.9847	225.8690	256.9107
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.9810
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	178.7926	201.1604
16	A	1	M	211.0537	105.8942	128.4316	160.4078	182.0009
17	B	1	M	209.6373	129.5539	154.0173	193.6574	221.4859
18	B	3	M	198.5871	116.3291	138.8718	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	199.8970
20	B	2	F	214.3518	121.6336	144.4541	182.0825	205.8866

Complete dataset with true value

Simulated complete dataset

- Visualisation of the first 20 IDs ...

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	134.6131	158.7102	201.8993	229.4077
2	B	3	F	205.4844	128.5105	152.9126	191.3709	219.0251
3	B	3	M	220.8016	130.4923	156.0304	195.2262	221.5212
4	B	2	M	208.1906	147.7882	176.9052	221.7698	252.6078
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	160.9798	202.3034	230.9175
7	A	2	M	215.7308	136.5162	165.6273	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.7470	212.1675
9	A	2	F	221.9858	152.7514	182.1968	228.9192	260.2738
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.2210	225.7538
12	A	1	M	209.8337	149.9582	180.9847	225.8690	256.9107
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.9810
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	178.7926	201.1604
16	A	1	M	211.0537	105.8942	128.4316	160.4078	182.0009
17	B	1	M	209.6373	129.5539	154.0173	193.6574	221.4859
18	B	3	M	198.5871	116.3291	138.8718	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	199.8970
20	B	2	F	214.3518	121.6336	144.4541	182.0825	205.8866

Complete dataset with true value

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?
7	A	2	M	215.7308	136.5162	?	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.747	?
9	A	2	F	221.9858	152.7514	?	?	?
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.221	?
12	A	1	M	209.8337	149.9582	?	?	?
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.981
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	?	201.1604
16	A	1	M	211.0537	105.8942	128.4316	?	182.0009
17	B	1	M	209.6373	129.5539	154.0173	?	?
18	B	3	M	198.5871	116.3291	?	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	?
20	B	2	F	214.3518	121.6336	144.4541	182.0825	?

Complete dataset in reality

Simulated complete dataset

- Visualisation of the first 20 IDs …

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	134.6131	158.7102	201.8993	229.4077
2	B	3	F	205.4844	128.5105	152.9126	191.3709	219.0251
3	B	3	M	220.8016	130.4923	156.0304	195.2262	221.5212
4	B	2	M	208.1906	147.7882	176.9052	221.7698	252.6078
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	160.9798	202.3034	230.9175
7	A	2	M	215.7308	136.5162	165.6273	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.7470	212.1675
9	A	2	F	221.9858	152.7514	182.1968	228.9192	260.2738
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.2210	225.7538
12	A	1	M	209.8337	149.9582	180.9847	225.8690	256.9107
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.9810
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	178.7926	201.1604
16	A	1	M	211.0537	105.8942	128.4316	160.4078	182.0009
17	B	1	M	209.6373	129.5539	154.0173	193.6574	221.4859
18	B	3	M	198.5871	116.3291	138.8718	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	199.8970
20	B	2	F	214.3518	121.6336	144.4541	182.0825	205.8866

Complete dataset with true value

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?
7	A	2	M	215.7308	136.5162	?	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.747	?
9	A	2	F	221.9858	152.7514	?	?	?
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.221	?
12	A	1	M	209.8337	149.9582	?	?	?
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.981
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	?	201.1604
16	A	1	M	211.0537	105.8942	128.4316	?	182.0009
17	B	1	M	209.6373	129.5539	154.0173	?	?
18	B	3	M	198.5871	116.3291	?	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	?
20	B	2	F	214.3518	121.6336	144.4541	182.0825	?

Complete dataset in reality

- How to handle this common situation ?
- Ignoring missing values ?
- Replacement by specific values ? Like LOCF by ex. ?

Simulate dataset : DESCRIPTIVES

- Simulate missing data
  - Completely observed : ID, x, method, educ, sex
  - Missing data for y_0 : 4 cases missing (2.7%)
  - Missing data for y_1 : 51 cases missing (34.0%)
  - Missing data for y_2 : 74 cases missing (49,3%)
  - Missing data for y_3 : 76 cases missing (50,7%)

Simulate dataset : DESCRIPTIVES

- Simulate missing data
  - Completely observed : ID, x, method, educ, sex
  - Missing data for y_0 : 4 cases missing (2.7%)
  - Missing data for y_1 : 51 cases missing (34.0%)
  - Missing data for y_2 : 74 cases missing (49,3%)
  - Missing data for y_3 : 76 cases missing (50,7%)

Variables	n	Mean	True Mean	SD	True SD	95% CI for Mean
x	150	212.76	212.76	9.54	9.54	[211.22 - 214.3]
y_0	146	130.18	130.42	10.94	10.93	[128.38 - 131.98]
y_1	99	155.85	156.52	12.34	13.26	[153.38 - 158.32]
y_2	76	**190.94**	**195.54**	14.34	16.44	**[187.64 - 194.24]**
y_3	74	**216.20**	**221.67**	15.66	18.75	**[212.55 - 219.85]**

What is missing data ?

❑ LACK OF RESPONSE

- Subjects refuse to participate or do not show up
- Subjects drop out of a study
- Subjects cannot or refuse to answer specific questions
- Subjects give « Don't kwow » answers
- Written answer is unreadable
- …

❑ IMPORTANT QUESTION

⇨ **DOES AN UNDERLYING TRUE VALUE EXIST ?**

Why are missing data a problem ?

- Most data analysis procedures and software were not designed to handle missing data
  - Solution ? … Ignoring missingness ?
  - Still today, the default option in too many softwares

Why are missing data a problem ?

- Most data analysis procedures and software were not designed to handle missing data
  - Solution ? … Ignoring missingness ?
  - Still today, the default option in too many softwares

- Ignoring missing data or imputations lends an appearance of completeness, but may lead to serious problem
  - Example (simulated dataset) : Comparing the means of 2 independant variabes (x and y_3) using a classical t-test

Why are missing data a problem ?

- Most data analysis procedures and software were not designed to handle missing data
  - Solution ? … Ignoring missingness ?
  - Still today, the default option in too many softwares

- Ignoring missing data or imputations lends an appearance of completeness, but may lead to serious problem
  - Example (simulated dataset) : Comparing the means of 2 independant variabes (x and y_3) using a classical t-test

Variables	n	t for 95% CI	P-value	Diff. y_3-x	Estimated power of test
Original	150	5.19	$4.8*10^{-7}$	9.54	0.999
Non - missing	74	1.73	0.09	10.94	0.415

Why are missing data a problem ?

- Most data analysis procedures and software were not designed to handle missing data
  - Solution ? … Ignoring missingness ?
  - Still today, the default option in too many softwares

- Ignoring missing data or imputations lends an appearance of completeness, but may lead to serious problem
  - Example (simulated dataset) : Comparing the means of 2 independant variabes (x and y_3) using a classical t-test

Variables	n	t for 95% CI	P-value	Diff. y_3-x	Estimated power of test
Original	150	5.19	$4.8*10^{-7}$	9.54	0.999
Non - missing	74	1.73	0.09	10.94	0.415

⇨ Inefficiency due to loss of information and sample size : **LOSS OF POWER**

Why are missing data a problem ?

- Biased results, depending on:
  - Systematic differences between responders and non-responders

	Reponders			Non-responders		
Variables	n	Mean	SD	n	Mean	SD
y_2	76	190.94	14.34	74	200.26	17.20
y_3	74	216.20	15.66	76	227.00	20.03

Why are missing data a problem ?

- Biased results, depending on:
  - Systematic differences between responders and non-responders

	Reponders			Non-responders		
Variables	n	Mean	SD	n	Mean	SD
y_2	76	190.94	14.34	74	200.26	17.20
y_3	74	216.20	15.66	76	227.00	20.03

  - The proportion of missing data

Why are missing data a problem ?

- Biased results, depending on:
  - Systematic differences between responders and non-responders

	Reponders			Non-responders		
Variables	n	Mean	SD	n	Mean	SD
y_2	76	190.94	14.34	74	200.26	17.20
y_3	74	216.20	15.66	76	227.00	20.03

  - The proportion of missing data

In conclusion …

⇨ Inevitable loss of precision

⇨ Bias which depends on the choice of the proposed statistical model for inferences and how to take into account the missingness

Two moments for action

- **BEFORE AND DURING DATA COLLECTION**
  - Sampling design, data collection process, question answer process, determinants (sources) of non-response
  - Knowledge about causes of missingness : PREVENTION
- **AFTER DATA COLLECTION**
  - Statistical treatment

Two moments for action

❑ BEFORE AND DURING DATA COLLECTION

- Sampling design, data collection process, question answer process, determinants (sources) of non-response
- Knowledge about causes of missingness : PREVENTION

❑ AFTER DATA COLLECTION

- Statistical treatment

⇨ Available case,
Frequentists methods (Weighting)
Likelihood-based
Imputation
Pattern-mixture model …

What determine a missing data ?

- Three potential determinants (sources) of missing data can be distinguished (Huisman & Van Der Zouwen, 1999)

What determine a missing data ?

- Three potential determinants (sources) of missing data can be distinguished (Huisman & Van Der Zouwen, 1999)

1. The researcher: Influences response behavior through

What determine a missing data ?

- Three potential determinants (sources) of missing data can be distinguished (Huisman & Van Der Zouwen, 1999)

1. The researcher: Influences response behavior through
   - The design of the study: Mode of data collection

What determine a missing data ?

- **Three potential determinants (sources) of missing data can be distinguished (Huisman & Van Der Zouwen, 1999)**

1. The researcher: Influences response behavior through
   - The design of the study: Mode of data collection
   - The design of the questionnaire: Structure, wording, tasks, format, don't know, …

What determine a missing data ?

- **Three potential determinants (sources) of missing data can be distinguished (Huisman & Van Der Zouwen, 1999)**

1. The researcher: Influences response behavior through
   - The design of the study: Mode of data collection
   - The design of the questionnaire: Structure, wording, tasks, format, don't know, …
2. The respondent
3. The interviewer

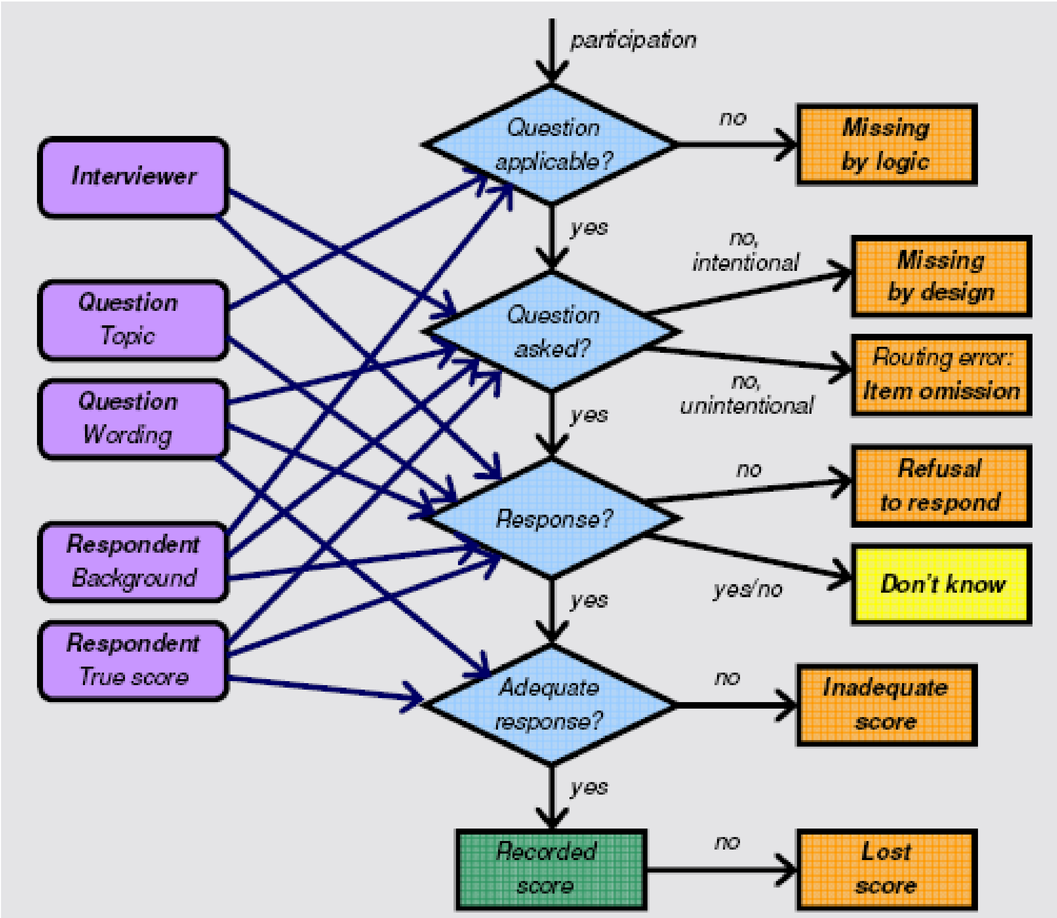
What determine a missing data ?

- **Three potential determinants (sources) of missing data can be distinguished (Huisman & Van Der Zouwen, 1999)**
  1. The researcher: Influences response behavior through
     - The design of the study: Mode of data collection
     - The design of the questionnaire: Structure, wording, tasks, format, don't know, …
  2. The respondent
  3. The interviewer

- **These are the source of errors (Groves, 1989)**
  - Mode of data collection, questionnaire, respondent, interviewer
  - Also in data processing, answers can be lost

Anticipating the missing data a priori

- **Step in data collection, question – answer and editing process**



- Diamonds: Step and decisions
- Ovals: Factors that affect decisions
- Boxes: Results with respect to non –response

⇩

Lead to different types of non -responses

Prevention

Problem / failure	*Prevention*
	Mode of data collection
Self-administered questionnaire generate nonresponse	Pretest layout and design, use interviewer, computer-assisted data collection
Internet questionnaire: representative?	Invite respondents, collect enough covariates
	Questionnaire
Layout: branching, length, item position	Pretest layout
Question topic: threat, sensitive	Use interviewers or not? Special formats
Question structure: wording, instruction, format, response categories (include *DK*?)	Pretest, expert reviews, use interviewers
Question difficulty: cognitive task	Keep respondent motivated, special formats
	Respondent
Skip, refuse, not be able, not understand	Look closely at *question-answer process*
Attributes: *correlates*	Special attention, use interviewers, questionnaire layout
	Interviewer
Fail to ask, record, probe	Interviewer training, computer-assisted, supervision
	Other
Data processing: entry, coding, editing	Computer-assisted data collection
Institutional requirements and policies: providing answers is voluntary	Instructions for interviewers: no probing, offer no-opinion options

THE COMPREHENSION OF MISSINGNESS

A typology of missing data

- Data are ***COMPLETELY*** *missing*
  - A sampled unit is not observed and the entire data collection fails

    Example: A subject finally refuses to participate to the study

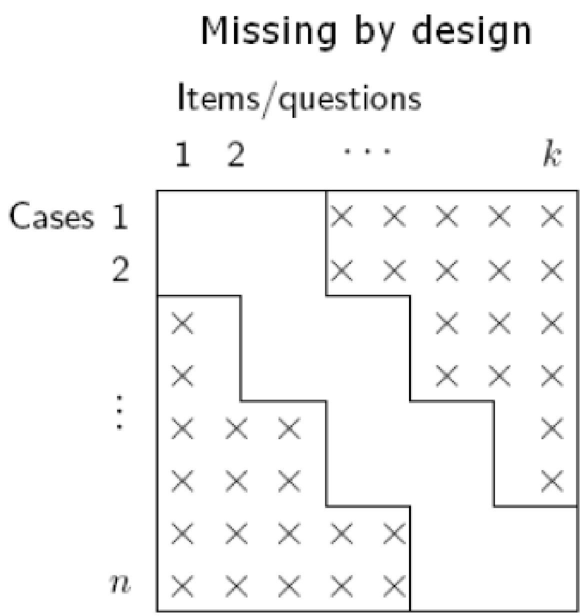
A typology of missing data

- Data are ***COMPLETELY*** *missing*
  - A sampled unit is not observed and the entire data collection fails

    Example: A subject finally refuses to participate to the study
- Data are ***PARTIALLY*** *missing*

A typology of missing data

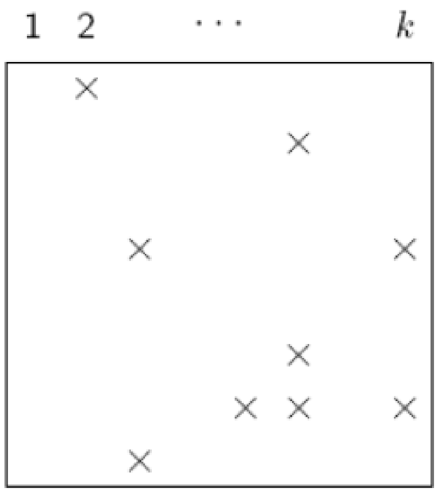
- Data are ***COMPLETELY*** *missing*
  - A sampled unit is not observed and the entire data collection fails
    Example: A subject finally refuses to participate to the study
- Data are ***PARTIALLY*** *missing*
  - Missing by design: Missingness created by the researcher, e.g., not-applicable items, incomplete designs



A typology of missing data

- Data are ***COMPLETELY*** *missing*
  - A sampled unit is not observed and the entire data collection fails
    Example: A subject finally refuses to participate to the study
- Data are ***PARTIALLY*** *missing*
  - Missing by design: Missingness created by the researcher, e.g., not-applicable items, incomplete designs
  - Item non-response: Missingness on individual variables or items, e.g., skipped items, inadequate responses, information not recorded or lost

Item nonresponse

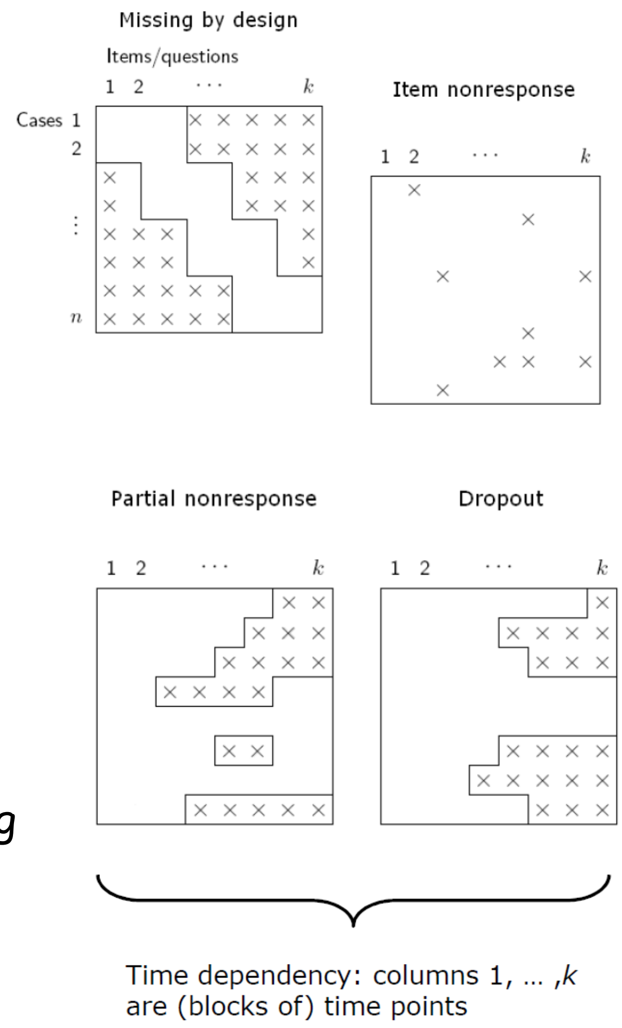


A typology of missing data

- Data are *COMPLETELY missing*
  - A sampled unit is not observed and the entire data collection fails

    Example: A subject finally refuses to participate to the study
- Data are *PARTIALLY missing*
  - Missing by design: Missingness created by the researcher, e.g., not-applicable items, incomplete designs
  - Item non-response: Missingness on individual variables or items, e.g., skipped items, inadequate responses, information not recorded or lost
  - Partial or wave non-response: Missingness depending on time/time points, e.g., dropout, attrition, missing baseline, break-off during interview

    Example: a subject involved in the study but does not respond to all questions



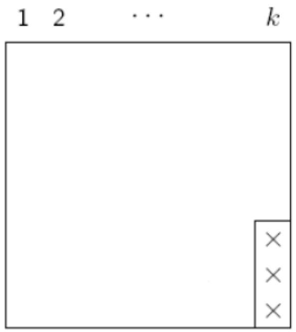
Patterns of missingness

- Data sets are assumed to be arranged in matrix form
- Different types show different patterns
- Important classes of overall missing data patterns distinguished :

Patterns of missingness

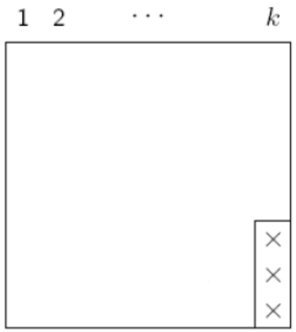
- Data sets are assumed to be arranged in matrix form
- Different types show different patterns
- Important classes of overall missing data patterns distinguished :
  - **Univariate pattern**

    Missing value occur on one item or group of itemsthat are either entirely observed or missing



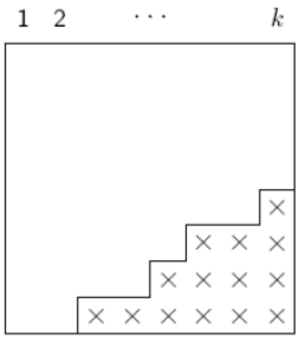
Patterns of missingness

- Data sets are assumed to be arranged in matrix form
- Different types show different patterns
- Important classes of overall missing data patterns distinguished :



- **Univariate pattern**

  Missing value occur on one item or group of itemsthat are either entirely observed or missing



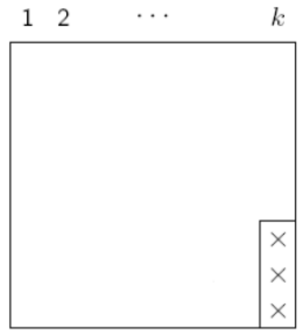
- **Monotone pattern**

  Items are ordered such that if item p is missing items p + 1, …, k are also missing

  ⇨ *frequently encountered in longitudinal studies*

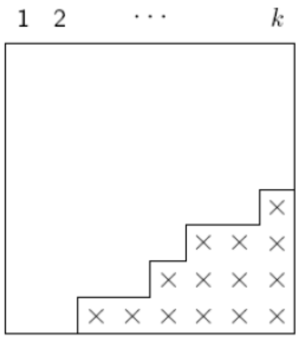
Patterns of missingness

- Data sets are assumed to be arranged in matrix form
- Different types show different patterns
- Important classes of overall missing data patterns distinguished :



- **Univariate pattern**

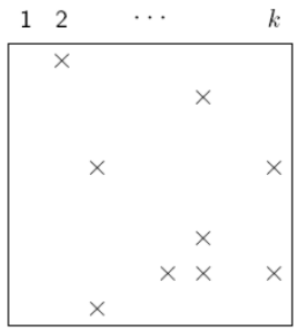
  Missing value occur on one item or group of itemsthat are either entirely observed or missing



- **Monotone pattern**

  Items are ordered such that if item p is missing items p + 1, …, k are also missing

  ⇨ *frequently encountered in longitudinal studies*



- **Arbitrary pattern**

  Random scatter of missing data

Standard theory

- Notations
  - Consider data set in matrix form

- Visualization on simulated datasets (n = 6 subjects)

	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?

Standard theory

- Notations
  - Consider data set in matrix form

  ⇨ **Y matrix of outcomes** with Y_i = outcomes for subject i at all occasions
  Y could be a vector (transversal study) or a matrix (longitudinal study)

- Visualization on simulated datasets (n = 6 subjects)

Y

	method	educ	sex	x	y0	y1	y2	y3	
1	A	2	F	230.2322	?	158.7102	?	229.4077	
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251	
3	B	3	M	220.8016	130.4923	?	?	221.5212	Y_3
4	B	2	M	208.1906	147.7882	176.9052	?	?	
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965	
6	A	2	M	201.9489	134.4354	?	202.3034	?	

Standard theory

- Notations
  - Consider data set in matrix form

    ➡ **Y matrix of outcomes** with Y_i = outcomes for subject i at all occasions
    Y could be a vector (transversal study) or a matrix (longitudinal study)

    ➡ **X matrix of covariates** with X_i = covariates for subject i
    (assume completely observed)

- Visualization on simulated datasets (n = 6 subjects)

	X				Y				
	method	educ	sex	x	y0	y1	y2	y3	
1	A	2	F	230.2322	?	158.7102	?	229.4077	
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251	
3	B	3	M	220.8016	130.4923	?	?	221.5212	Y_3
4	B	2	M	208.1906	147.7882	176.9052	?	?	
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965	
6	A	2	M	201.9489	134.4354	?	202.3034	?	

Standard theory

❑ Notations

- Consider data set in matrix form

  ⇨ **Y matrix of outcomes** with Y_i = outcomes for subject i at all occasions
  Y could be a vector (transversal study) or a matrix (longitudinal study)

  ⇨ **X matrix of covariates** with X_i = covariates for subject i
  (assume completely observed)

  ⇨ **R matrix of missingness** with R_i = binary variables indicating wether each element of Y_i is observed (1) or missing (0)

- Remark : If the only kind of missing data is DROPOUT, the missingness can be reduce to a D vector where D_i is the time of last measurement

❑ Visualization on simulated datasets (n = 6 subjects)

X | Y

	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?

Y_3

⇨

R

	method	educ	sex	x	y0	y1	y2	y3
1	1	1	1	1	0	1	0	1
2	1	1	1	1	1	1	0	1
3	1	1	1	1	1	0	0	1
4	1	1	1	1	1	1	0	0
5	1	1	1	1	1	1	1	1
6	1	1	1	1	1	0	1	0

$D_4=2$

Standard theory

- ***The model of measurement***
  - $P(Y_i|X_i,\theta)$ = some distribution

⇨ Characteristics of θ ?

Standard theory

- ***The model of measurement***
  - $P(Y_i|X_i,\theta)$ = some distribution

Characteristics of θ ?

- θ : Population parameters of interest
- It could be a scalar or a vector
- It could correspond to :
  - Effects of covariate on response
  - Difference in mean response at final occasion
- θ applies to the entire population of subjects

Standard theory

- ***The model of measurement***
  - $P(Y_i|X_i,\theta)$ = some distribution

⇨ Characteristics of θ ?

- θ : Population parameters of interest
- It could be a scalar or a vector
- It could correspond to :
  - Effects of covariate on response
  - Difference in mean response at final occasion
- θ applies to the entire population of subjects

- ***The distribution of missingness [DOM]***
  - Introduced by Rubin (1976, Biometrika), sometimes called the « missingness mechanism (or process) », **to clarify the conditions under which it may be ignored**
  - $P(R_i|X_i,Y_i,\Phi)$ = some distribution

Application on the simulated data set (n = 150)

- ***Remember …***

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?
7	A	2	M	215.7308	136.5162	?	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.747	?
9	A	2	F	221.9858	152.7514	?	?	?
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.221	?
12	A	1	M	209.8337	149.9582	?	?	?
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.981
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	?	201.1604
16	A	1	M	211.0537	105.8942	128.4316	?	182.0009
17	B	1	M	209.6373	129.5539	154.0173	?	?
18	B	3	M	198.5871	116.3291	?	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	?
20	B	2	F	214.3518	121.6336	144.4541	182.0825	?

Application on the simulated data set (n = 150)

- ***Remember …***

ID	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?
7	A	2	M	215.7308	136.5162	?	203.9909	233.3036
8	B	2	F	225.6780	126.5309	148.8798	188.747	?
9	A	2	F	221.9858	152.7514	?	?	?
10	A	1	F	204.5258	128.0827	155.5105	193.5588	217.6615
11	B	2	M	233.4815	131.6936	158.0804	198.221	?
12	A	1	M	209.8337	149.9582	?	?	?
13	A	1	M	199.7724	127.6674	154.7696	191.8104	215.981
14	B	2	F	222.1966	131.4658	156.7415	196.2984	222.9952
15	A	2	M	218.0227	118.3037	143.3665	?	201.1604
16	A	1	M	211.0537	105.8942	128.4316	?	182.0009
17	B	1	M	209.6373	129.5539	154.0173	?	?
18	B	3	M	198.5871	116.3291	?	174.8252	197.4983
19	A	1	M	227.8085	116.7549	139.1389	174.5858	?
20	B	2	F	214.3518	121.6336	144.4541	182.0825	?

- ***Standard descriptive table***

  Frequencies table of the different patterns observed ⇨

Measurement occasion (y)					
0	1	2	3	Number	%
Completers					
1	1	1	1	30	20.00
Dropouts / Monotone pattern					
1	1	0	0	25	16.67
1	1	1	0	23	15.31
1	0	0	0	15	10.00
0	0	0	0	1	0.67
Non-monotone pattern					
1	1	0	1	19	12.67
1	0	0	1	12	8.00
1	0	1	1	12	8.00
1	0	1	0	10	6.67
0	0	1	0	1	0.67
0	1	0	0	1	0.67
0	1	0	1	1	0.67

0: Observed 1: Missing

Classification of missingness mechanisms (Part I)

Based on Rubin (1976), Little & Rubin (1987) and Little (1995), ***4 missingness processes*** *were clearly defined :*

- ***Missing Completely at Random (MCAR)***

Classification of missingness mechanisms (Part I)

*Based on Rubin (1976), Little & Rubin (1987) and Little (1995), **4 missingness processes** were clearly defined :*

- ***Missing Completely at Random (MCAR)***
  - DOM does not depend on covariates or outcomes
  - $P(R_i|X_i,Y_i,\Phi) = P(R_i|\Phi)$
  - … means that the probability of missing information is unrelated to any characteristics of the subject at all

Classification of missingness mechanisms (Part I)

*Based on Rubin (1976), Little & Rubin (1987) and Little (1995), **4 missingness processes** were clearly defined :*

❑ *Missing Completely at Random (MCAR)*

- DOM does not depend on covariates or outcomes
- $P(R_i|X_i,Y_i,\Phi) = P(R_i|\Phi)$
- … means that the probability of missing information is unrelated to any characteristics of the subject at all

⇨ **Responders are representative (sub)sample of the population**

Classification of missingness mechanisms (Part I)

*Based on Rubin (1976), Little & Rubin (1987) and Little (1995), **4 missingness processes** were clearly defined :*

- ***Missing Completely at Random (MCAR)***
  - DOM does not depend on covariates or outcomes
  - $P(R_i|X_i,Y_i,\Phi) = P(R_i|\Phi)$
  - … means that the probability of missing information is unrelated to any characteristics of the subject at all

    ⇨ **Responders are representative (sub)sample of the population**

  - **Consequences on subsequent statisitical analysis**

    ⇨ **Loss of precision (power)**

    ⇨ No bias

Classification of missingness mechanisms (Part II)

- ***Covariate-dependent missingness (CD)***
  - DOM may possibly depend on covariates but not outcomes
  - $P(R_i|X_i,Y_i,\Phi) = P(R_i|x_i,\Phi)$
  - **In the case of dropouts**, CD means that the probability of missing information may be related to covariates but its unrelated to outcomes at any time

Classification of missingness mechanisms (Part II)

- ***Covariate-dependent missingness (CD)***
  - DOM may possibly depend on covariates but not outcomes
  - $P(R_i|X_i,Y_i,\Phi) = P(R_i|x_i,\Phi)$
  - **In the case of dropouts**, CD means that the probability of missing information may be related to covariates but its unrelated to outcomes at any time
- ***Missing at random (MAR)***
  - DOM may depend on covariates and observed outcomes
  - $P(R_i|X_i,Y_i,\Phi) = P(R_i|y_{i(obs)},x_i,\Phi)$
  - In the case of dropouts, **MAR means that the probability of missing information may be related to covariates and to pre-dropout responses**
  - **In longitudinal studies:** *Missingness may depend on previous measurements but not on actual and future measurements*

Classification of missingness mechanisms (Part III)

- Consequences of MAR on subsequent statistical analyses

⇨ MCAR ⊂ CD ⊂ MAR

⇨ Non-response can be predicted from observed data

⇨ Loss of precision (power)

⇨ No bias with appropriate statistical methods

Classification of missingness mechanisms (Part IV)

Missing not at random (MNAR)

- DOM still depends on $y_{i(miss)}$ even after any dependence on x_i and $y_{i(obs)}$ has been accounted for
- $P(R_i|X_i,Y_i,\Phi) = P(R_i|y_{i(obs)},y_{i(miss)},x_i,\Phi)$
- In a monotone pattern, MNAR means that the probability of dropout is related to responses at the time of dropout and possibility afterward

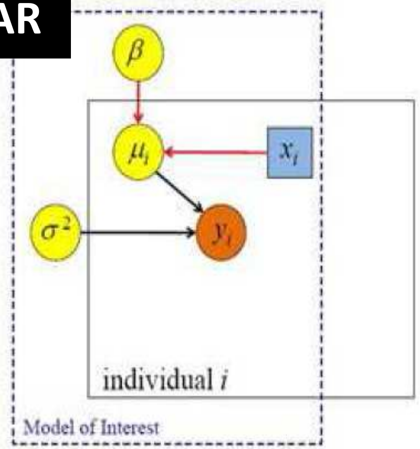
Classification of missingness mechanisms (Part IV)

Missing not at random (MNAR)

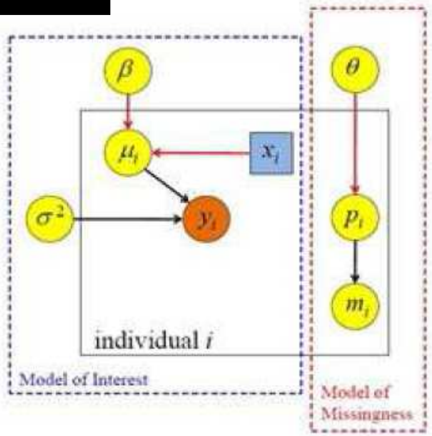
- DOM still depends on $y_{i(miss)}$ even after any dependence on x_i and $y_{i(obs)}$ has been accounted for
- $P(R_i|X_i,Y_i,\Phi) = P(R_i|y_{i(obs)},y_{i(miss)},x_i,\Phi)$
- In a monotone pattern, MNAR means that the probability of dropout is related to responses at the time of dropout and possibility afterward
- **Consequences on subsequent statistical analyses**
  - ⇨ **Loss of precision (power)**
  - ⇨ **Systematic bias due to systematic differences**
  - ⇨ **Requires advanced modeling of missing and observed data**

Description of the missingness mechanisms in a Bayes procedure (Bugs diagram)

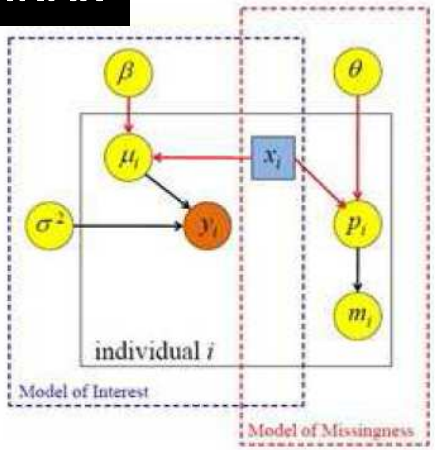
MCAR



MAR



MNAR



❑ Model of interest

MCAR
- $y_i \sim N(\mu_i;\sigma^2)$
- $\mu_i = x_i\beta$
- β ~ fully specified by an a priori distribution

MAR
- $y_i \sim N(\mu_i;\sigma^2)$
- $\mu_i = x_i\beta$
- β ~ fully specified by an a priori distribution

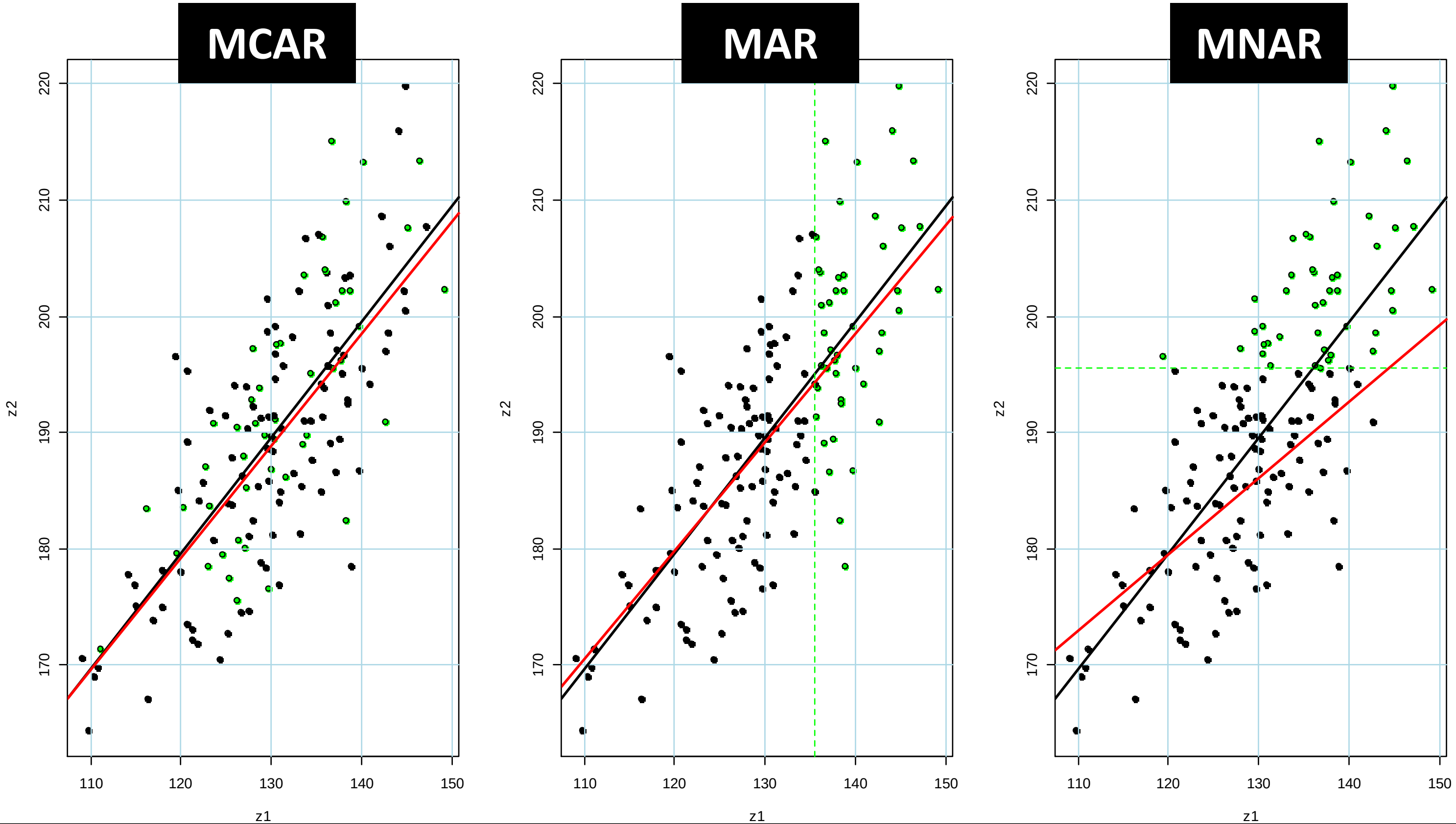
MNAR
- $y_i \sim N(\mu_i;\sigma^2)$
- $\mu_i = x_i\beta$
- β ~ fully specified by an a priori distribution

❑ Model of missingness

MAR
- $m_i \sim \text{Bernoulli}(p_i)$
- $\text{logit}(p_i) \sim \theta$

MNAR
- $m_i \sim \text{Bernoulli}(p_i)$
- $\text{logit}(p_i) \sim \theta\ (+x_i) + y_i$

Application on initial simulated dataset (N=150)



z2	n	Mean	SD	Intercept	Slope
Completely observed	150	189.84	11.08	59.92	0.99 (0.07)
MCAR	102	188.73	11.10	63.27	0.97 (0.09)
MAR	105	185.63	9.09	67.67	0.93 (0.11)
MNAR	105	184.30	7.57	100.43	0.66 (0.08)

Particularities of missingness mechanisms

- ***What can we tell from the data ?***

  - Because we observe x_i, r_i, and $y_{i(obs)}$, it is often possible to reject MCAR and CD in favor of MAR

Particularities of missingness mechanisms

- ***What can we tell from the data ?***
  - Because we observe x_i, r_i, and $y_{i(obs)}$, it is often possible to reject MCAR and CD in favor of MAR
  - It is never possible to reject MAR in favor of MNAR on the base of observed data, because we cannot see $y_{i(miss)}$

Particularities of missingness mechanisms

What can we tell from the data ?

- Because we observe x_i, r_i, and $y_{i(obs)}$, it is often possible to reject MCAR and CD in favor of MAR
- It is never possible to reject MAR in favor of MNAR on the base of observed data, because we cannot see $y_{i(miss)}$

When may we ignore the missingness mechanism and not model it ?

- The conditions under which we may ignore the DOM **depending on the mode of inference for Φ** (frequentist, likelihood, Bayesian)

Particularities of missingness mechanisms

- ***What can we tell from the data ?***
  - Because we observe x_i, r_i, and $y_{i(obs)}$, it is often possible to reject MCAR and CD in favor of MAR
  - It is never possible to reject MAR in favor of MNAR on the base of observed data, because we cannot see $y_{i(miss)}$
- ***When may we ignore the missingness mechanism and not model it ?***
  - The conditions under which we may ignore the DOM **depending on the mode of inference for Φ** (frequentist, likelihood, Bayesian)

⇨ **FREQUENTIST** statistical procedures: Ignore the DOM only when the missing data are **<u>MCAR</u>** (OLS, GEE …)

Particularities of missingness mechanisms

- ***What can we tell from the data ?***
  - Because we observe x_i, r_i, and $y_{i(obs)}$, it is often possible to reject MCAR and CD in favor of MAR
  - It is never possible to reject MAR in favor of MNAR on the base of observed data, because we cannot see $y_{i(miss)}$

- ***When may we ignore the missingness mechanism and not model it ?***
  - The conditions under which we may ignore the DOM **depending on the mode of inference for** Φ (frequentist, likelihood, Bayesian)

⇨ **FREQUENTIST** statistical procedures: Ignore the DOM only when the missing data are **<u>MCAR</u>** (OLS, GEE …)

⇨ **LIKELIHOOD\ BAYES** procedures: Ignore the DOM only when the missing data are MAR (LMM, GLMM…)

Classification of missingness mechanisms (Part V)

- ***IGNORABLE or NOT IGNORABLE processus ?***

Classification of missingness mechanisms (Part V)

- *IGNORABLE or NOT IGNORABLE processus ?*

MCAR \MAR	⟺	IGNORABLE
MNAR	⟺	NOT IGNORABLE

[Little & Rubin,1987]

- *When do we have to model the missingness (DOM) ?*

Classification of missingness mechanisms (Part V)

- ***IGNORABLE or NOT IGNORABLE processus ?***

MCAR \MAR	⟺	IGNORABLE
MNAR	⟺	NOT IGNORABLE

[Little & Rubin,1987]

- ***When do we have to model the missingness (DOM) ?***

  - If we are doing **frequentist analyses** …

    ⇨ … we will have to model R_i, if the missingness **IS NOT MCAR**

Classification of missingness mechanisms (Part V)

- ***IGNORABLE or NOT IGNORABLE processus ?***

MCAR \MAR	⟺	IGNORABLE
MNAR	⟺	NOT IGNORABLE

[Little & Rubin,1987]

- ***When do we have to model the missingness (DOM) ?***

  - If we are doing **frequentist analyses** …

    ⇨ … we will have to model R_i, if the missingness **IS NOT MCAR**

  - If we are doing **Likelihood or Bayesian analyses** …

    ⇨ … we will have to model R_i, **ONLY IF** we believe that missingness **is MNAR**

How detect the mechanism of missingness (DOM) ?

1. In some (too rarely) cases, DOM can be clearly discernable …

How detect the mechanism of missingness (DOM) ?

1 In some (too rarely) cases, DOM can be clearly discernable …

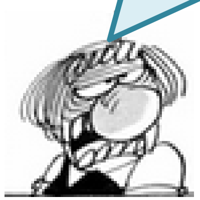
" Is it plausible for the data to be MCAR ? If so, then I would say there is evidence that the data is MCAR."

How detect the mechanism of missingness (DOM) ?

1 In some (too rarely) cases, DOM can be clearly discernable …

" Is it plausible for the data to be MCAR ? If so, then I would say there is evidence that the data is MCAR."

"The measurement is too expensive, she is only done on a sub-sample"



How detect the mechanism of missingness (DOM) ?

1 In some (too rarely) cases, DOM can be clearly discernable …

" Is it plausible for the data to be MCAR ? If so, then I would say there is evidence that the data is MCAR."

"The measurement is too expensive, she is only done on a sub-sample"

MCAR !

How detect the mechanism of missingness (DOM) ?

1 In some (too rarely) cases, DOM can be clearly discernable …

" Is it plausible for the data to be MCAR ? If so, then I would say there is evidence that the data is MCAR."

"The measurement is too expensive, she is only done on a sub-sample"

MCAR !

"A subject misses an evaluation because of a transport strike"

How detect the mechanism of missingness (DOM) ?

1 In some (too rarely) cases, DOM can be clearly discernable …

" Is it plausible for the data to be MCAR ? If so, then I would say there is evidence that the data is MCAR."

"The measurement is too expensive, she is only done on a sub-sample"

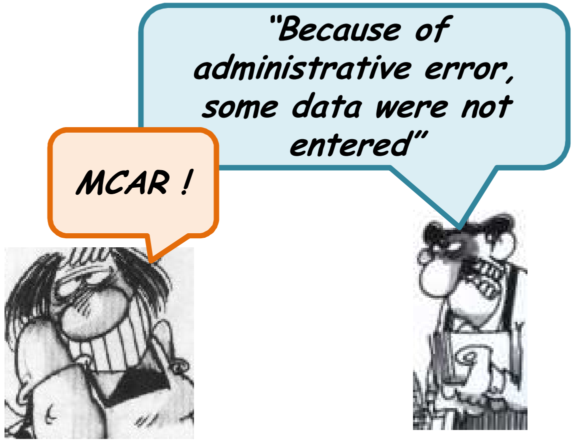
MCAR !

"A subject misses an evaluation because of a transport strike"

MCAR !

Examples of discernables missingness mechanisms

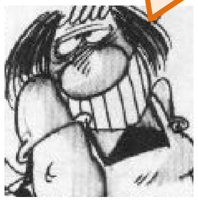
Examples of discernables missingness mechanisms



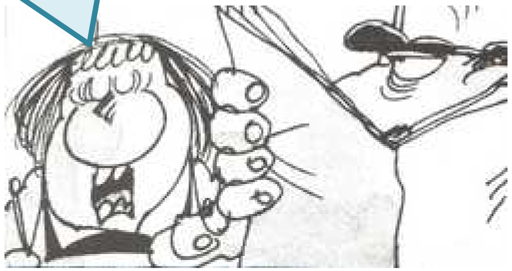
Examples of discernables missingness mechanisms

"Because of administrative error, some data were not entered"

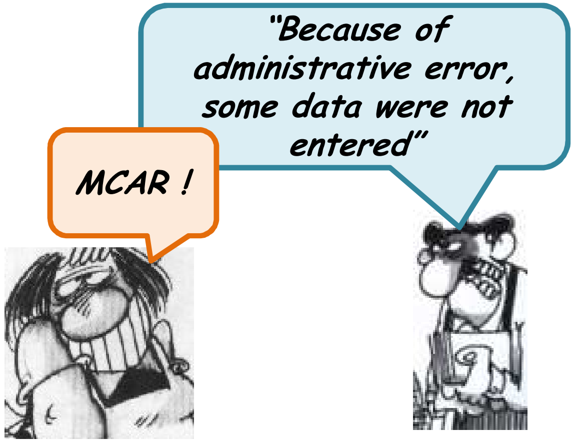
MCAR !

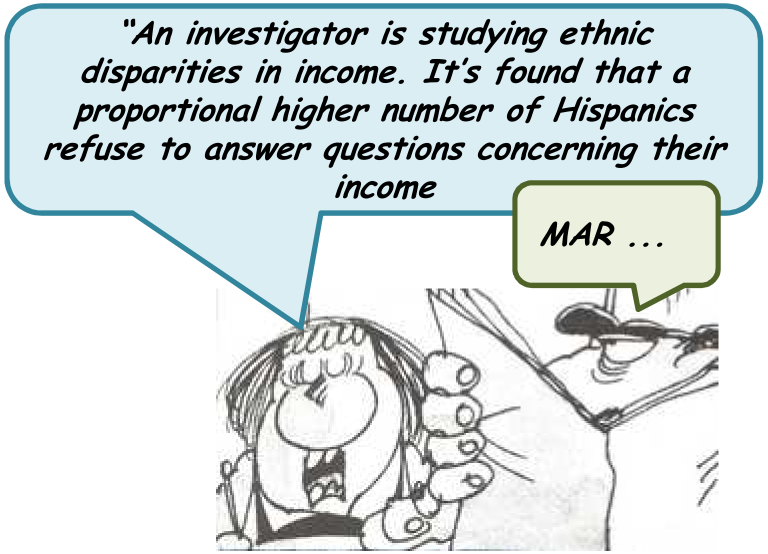


"An investigator is studying ethnic disparities in income. It's found that a proportional higher number of Hispanics refuse to answer questions concerning their income



Examples of discernables missingness mechanisms

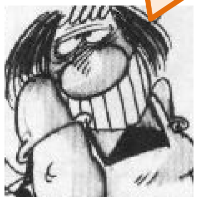




Examples of discernables missingness mechanisms

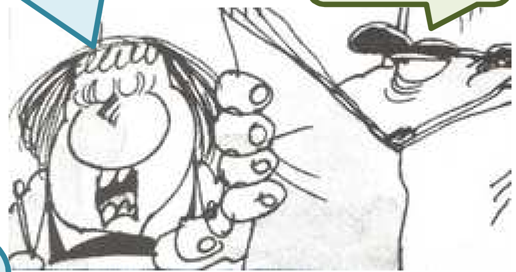
"Because of administrative error, some data were not entered"

MCAR !



"An investigator is studying ethnic disparities in income. It's found that a proportional higher number of Hispanics refuse to answer questions concerning their income

MAR ...

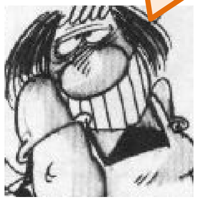


"An investigator is examining the effect of sleep on pain. Subjects are called daily and asked questions about last night's sleep and their pain today. Patients who are experiencing severe pain are more likely to not come to the phone leaving the data missing for that particular day

Examples of discernables missingness mechanisms

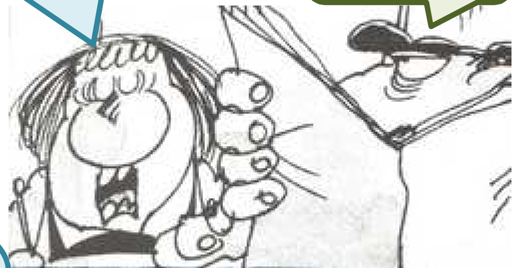
"Because of administrative error, some data were not entered"

MCAR !



"An investigator is studying ethnic disparities in income. It's found that a proportional higher number of Hispanics refuse to answer questions concerning their income

MAR ...



"An investigator is examining the effect of sleep on pain. Subjects are called daily and asked questions about last night's sleep and their pain today. Patients who are experiencing severe pain are more likely to not come to the phone leaving the data missing for that particular day

MNAR !

How detect the mechanism of missingness (DOM) ?

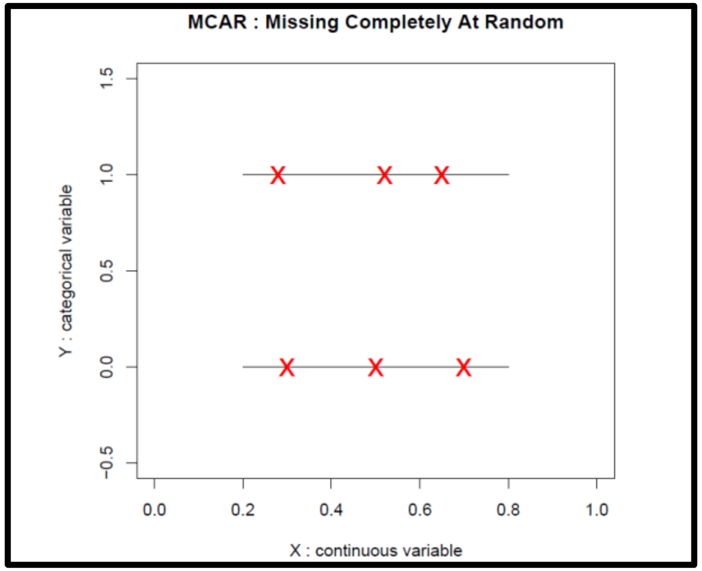
2 Graphical issues (descriptive techniques)

- Suppose that we want to study the DOM linked to a categorical variable Y
- Standard xyplot may help you to conclude

How detect the mechanism of missingness (DOM) ?

2 Graphical issues (descriptive techniques)

- Suppose that we want to study the DOM linked to a categorical variable Y
- Standard xyplot may help you to conclude

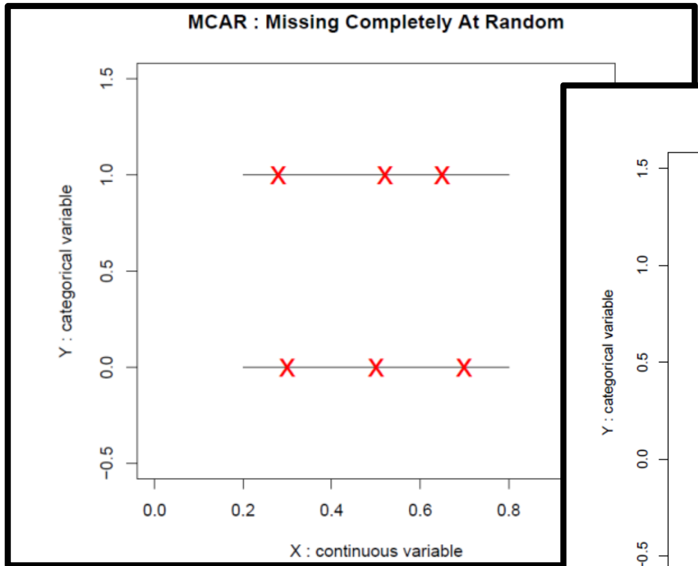


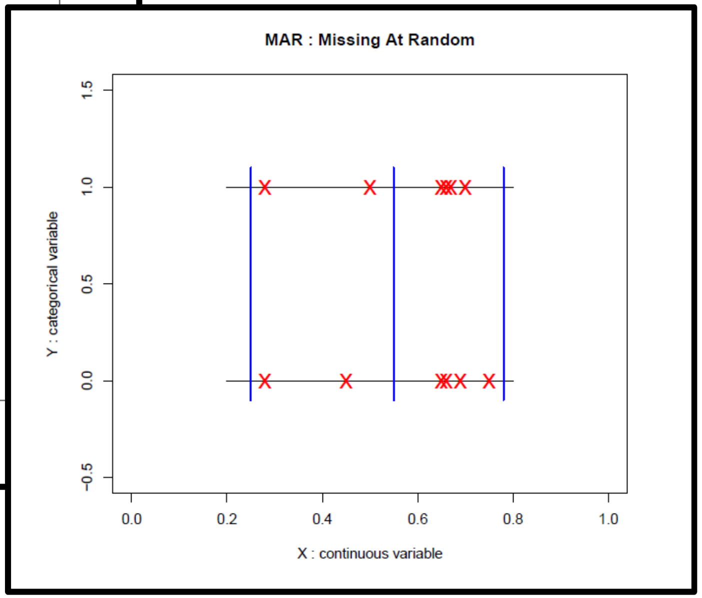
In this case, the continuous variable X seems to have no influence on the proportion of missingness within each group of Y…

How detect the mechanism of missingness (DOM) ?

2 Graphical issues (descriptive techniques)

- Suppose that we want to study the DOM linked to a categorical variable Y
- Standard xyplot may help you to conclude



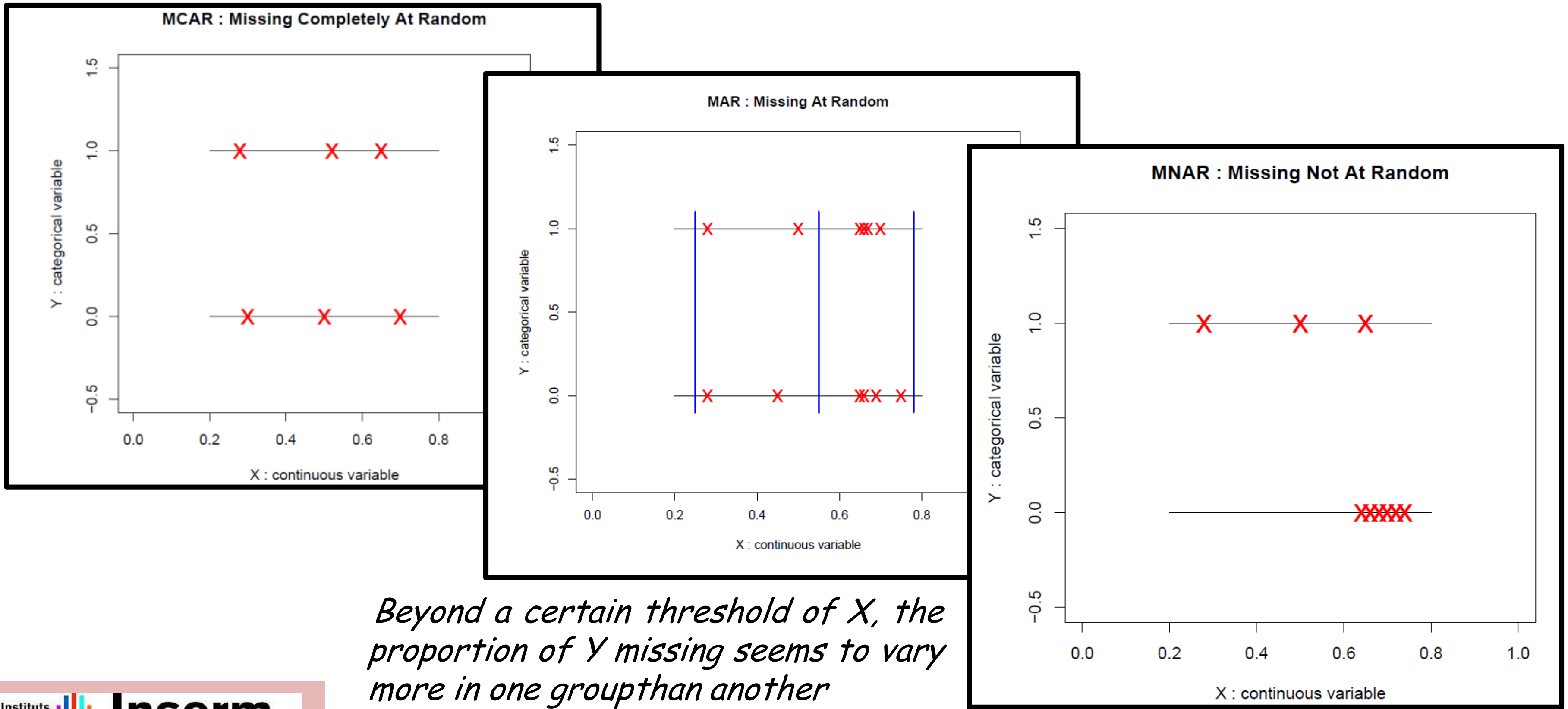


Beyond a certain threshold of X, the proportion of Y missing seems to vary

How detect the mechanism of missingness (DOM) ?

2 Graphical issues (descriptive techniques)

- Suppose that we want to study the DOM linked to a categorical variable Y
- Standard xyplot may help you to conclude



Beyond a certain threshold of X, the proportion of Y missing seems to vary more in one groupthan another

How detect the mechanism of missingness (DOM) ?

2 Graphical issues (descriptive techniques)

- … Searching for structural distribution of missingness with heatmap graphs

- In specific statistical software,

[R soft.]

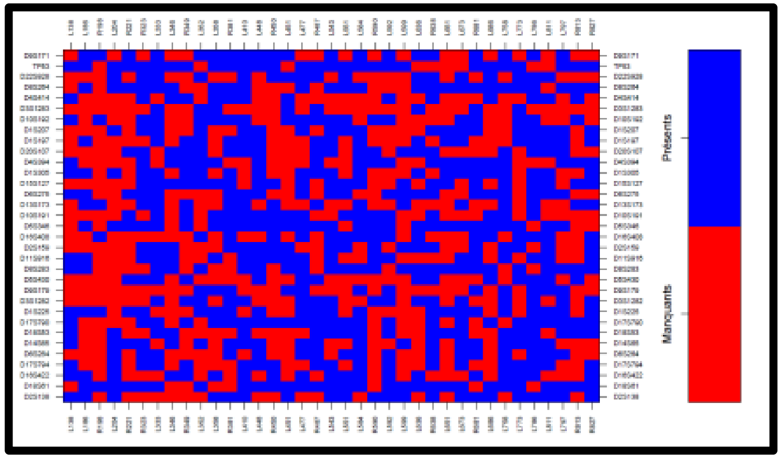
The package VIM (visualization and imputation of missing values) [Templ et al., 2011] is developed to explore and analyze the structure of missing data using graphical methods

[SPSS soft.]

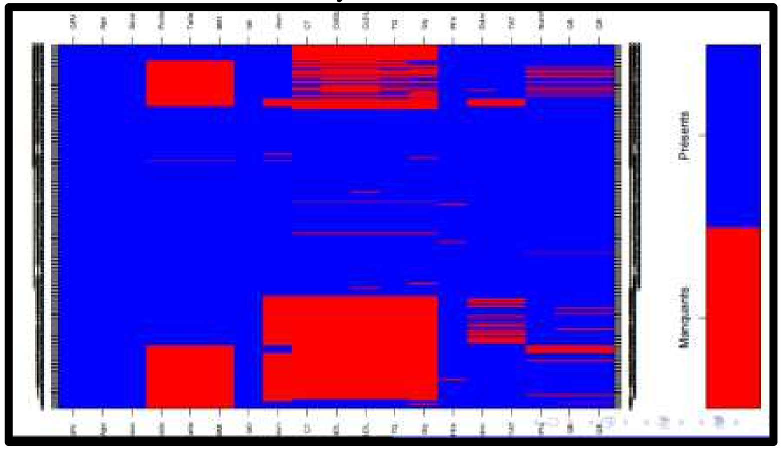
Procedure MVA

[…]

Uniform DOM



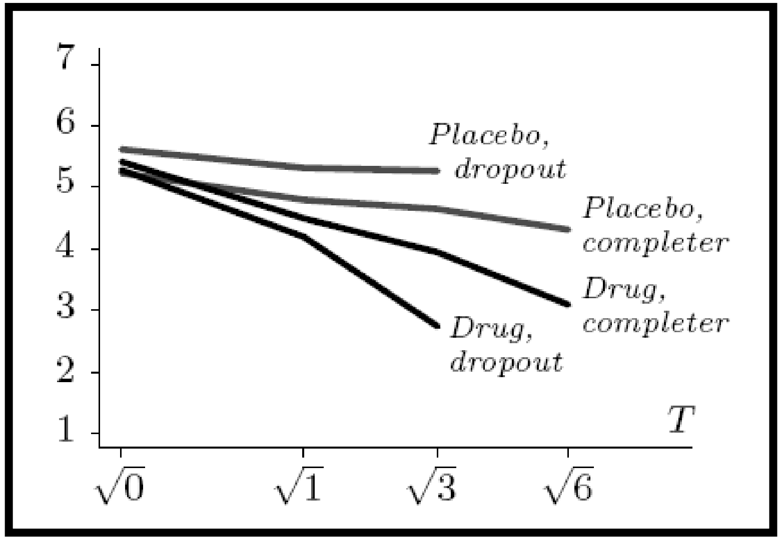
DOM with specific structure



How detect the mechanism of missingness (DOM) ?

2 Graphical issues

Plot of average response versus square root of week

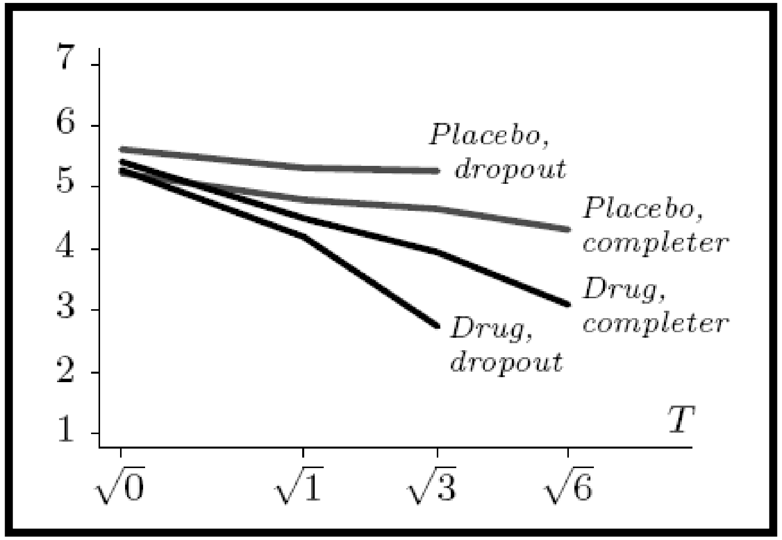


- An example based on a longitudinal study,
  - A randomized psychiatric trial
  - 312 patients received drug therapy for schizophrenia, 101 received placebo
  - Measurements at weeks 0, 1, 3, 6
  - Missing data primarily due to dropout
  - Outcome severity of illness from 1 to 7

How detect the mechanism of missingness (DOM) ?

2 Graphical issues

Plot of average response versus square root of week



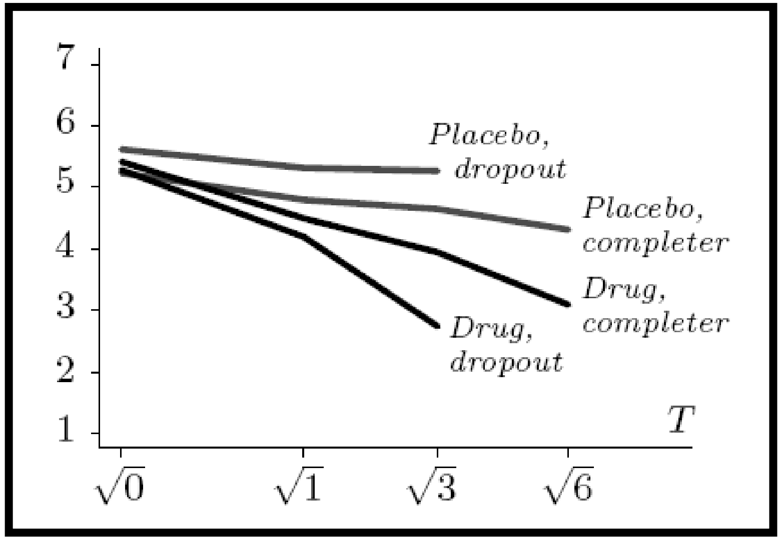
- An example based on a longitudinal study,
  - A randomized psychiatric trial
  - 312 patients received drug therapy for schizophrenia, 101 received placebo
  - Measurements at weeks 0, 1, 3, 6
  - Missing data primarily due to dropout
  - Outcome severity of illness from 1 to 7

- **Based on this plot**, we could conclude:

How detect the mechanism of missingness (DOM) ?

2 Graphical issues

Plot of average response versus square root of week



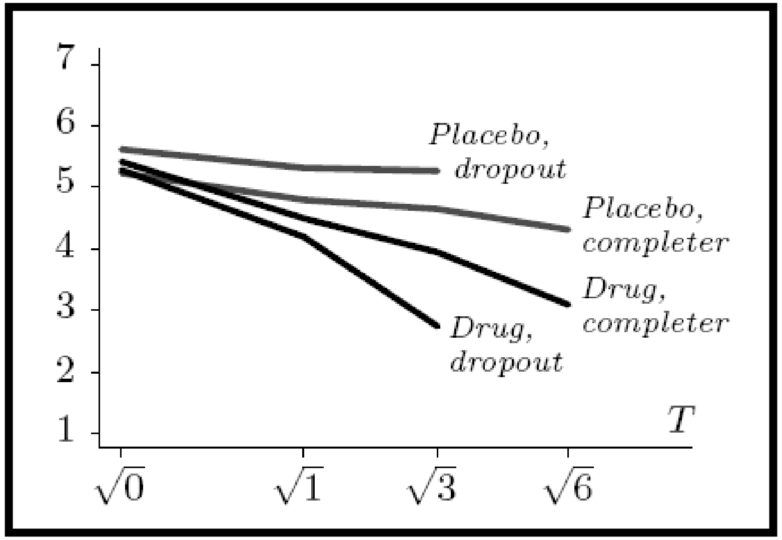
- An example based on a longitudinal study,
  - A randomized psychiatric trial
  - 312 patients received drug therapy for schizophrenia, 101 received placebo
  - Measurements at weeks 0, 1, 3, 6
  - Missing data primarily due to dropout
  - Outcome severity of illness from 1 to 7

- **Based on this plot**, we could conclude:
  - Dropout is not MCAR, because it operates differently in the treatment and control groups

How detect the mechanism of missingness (DOM) ?

2 Graphical issues

Plot of average response versus square root of week



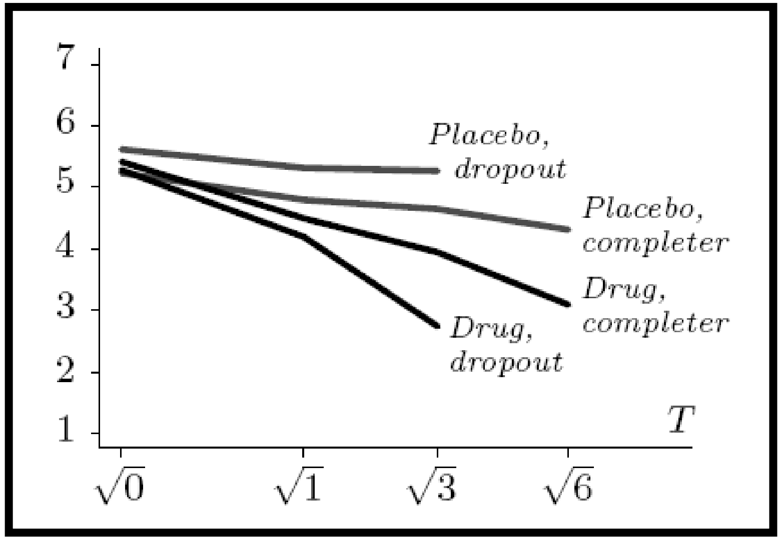
- An example based on a longitudinal study,
  - A randomized psychiatric trial
  - 312 patients received drug therapy for schizophrenia, 101 received placebo
  - Measurements at weeks 0, 1, 3, 6
  - Missing data primarily due to dropout
  - Outcome severity of illness from 1 to 7

- **Based on this plot**, we could conclude:
  - Dropout is not MCAR, because it operates differently in the treatment and control groups
  - Dropout is not merely CD, because completers and dropouts follow different (pre-dropout) trajectories

How detect the mechanism of missingness (DOM) ?

2 Graphical issues

Plot of average response versus square root of week



- An example based on a longitudinal study,
  - A randomized psychiatric trial
  - 312 patients received drug therapy for schizophrenia, 101 received placebo
  - Measurements at weeks 0, 1, 3, 6
  - Missing data primarily due to dropout
  - Outcome severity of illness from 1 to 7

- **Based on this plot**, we could conclude:

  - Dropout is not MCAR, because it operates differently in the treatment and control groups
  - Dropout is not merely CD, because completers and dropouts follow different (pre-dropout) trajectories
  - Dropout could be MAR or MNAR: it's impossible to tell

How detect the mechanism of missingness (DOM) ?

3 Assuming MAR

- MAR stay plausible in many settings
  - Example: In the « *missing by design cases* »,
    the mechanism of missingness is control by researcher/ experimenter
- MAR is ***only an assumption*** …
  - If the researcher does not control the mechanism or the mechanism is unknown
  - Find causes and correlates that confirm this assumption
  - Check: Follow-up of nonrepondents
- Violation of assumption
  - If there is no serious proof of non-randomness, erroneous assumption of MAR often has minor impact [Collins, 2001]

How detect the mechanism of missingness (DOM) ?

3 Checking MAR by testing

- Little's MCAR test

  - Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202.

  - Under the null hypothesis H_0, the DOM follows an MCAR process and the asymptotic distribution of the statistic is chi2

  - Presents many defaults

⇨ The test is too conservative with small sample

⇨ The test is more appropriate for continuous variables

How detect the mechanism of missingness (DOM) ?

3 Checking MAR by testing

- Little's MCAR test
  - Application to the original simulated dataset (n = 150) …

    ⇨ Using the LittleMCAR() function of the BaylorEdPsych package for R software, CHI.stat = 70.25, DF = 31, pvalue = $7.07*10^{-5}$

    ⇨ **Permit to conclude in favour of the MAR process, which is logic by construction of the Ys variables which depended on specific covariates by construction**

How detect the mechanism of missingness (DOM) ?

3 Checking MAR by testing

- Regression methods in **transversal studies**

  - Suppose that we want to know if the DOM linked to an Y variable follows a MAR process in a study of n subjects

  - Suppose that we dispose of n complete covariates, we can stock in a matrix plan X

➡ The modelisation of the binary vector of missingness R linked to Y gives us information about the process of interest:

$$\text{logit}(R_i) = X_i^{\dagger}\beta \text{ , } i = 1..n$$

➡ **The nullity of the vector of parameters β give us information in favour of an MCAR process**

How detect the mechanism of missingness (DOM) ?

3 Checking MAR by testing

- Regression methods in **longitudinal studies**
  - Dropouts are the processus of interest modelled by the time of last measurement D_i for patient i

$$\begin{cases} D_{ij} = 1 \text{ if visit j is the last one for patient i} \\ D_{ij} = 0 \text{ otherwise} \end{cases}$$

  - Remember the initial simulated dataset …

	X				Y			
	method	educ	sex	x	y0	y1	y2	y3
1	A	2	F	230.2322	?	158.7102	?	229.4077
2	B	3	F	205.4844	128.5105	152.9126	?	219.0251
3	B	3	M	220.8016	130.4923	?	?	221.5212
4	B	2	M	208.1906	147.7882	176.9052	?	?
5	B	2	F	215.2947	127.0137	151.5866	190.5087	216.0965
6	A	2	M	201.9489	134.4354	?	202.3034	?

How detect the mechanism of missingness (DOM) ?

3 Checking MAR by testing

- Regression methods in **longitudinal studies**

  - Dropout model by visit:

At visit 1:	$\text{logit}(D_{i1}) = X_i^{t}\beta_1$,	i= 1..n
At visit 2:	$\text{logit}(D_{i2}) = X_i^{t}\beta_2 + \eta_{20}Y_{i0}$,	
At visit 3:	$\text{logit}(D_{i3}) = X_i^{t}\beta_3 + \eta_{30}Y_{i0} + \eta_{31}Y_{i1}$	
At visit 4:	$\text{logit}(D_{i4}) = X_i^{t}\beta_4 + \eta_{40}Y_{i0} + \eta_{41}Y_{i1} + \eta_{42}Y_{i2}$	

How detect the mechanism of missingness (DOM) ?

3 Checking MAR by testing

- Regression methods in **longitudinal studies**

  - Dropout model by visit:

At visit 1:	$\text{logit}(D_{i1}) = X_i^{\dagger}\beta_1$,	i= 1..n
At visit 2:	$\text{logit}(D_{i2}) = X_i^{\dagger}\beta_2 + \eta_{20}Y_{i0}$,	
At visit 3:	$\text{logit}(D_{i3}) = X_i^{\dagger}\beta_3 + \eta_{30}Y_{i0} + \eta_{31}Y_{i1}$	
At visit 4:	$\text{logit}(D_{i4}) = X_i^{\dagger}\beta_4 + \eta_{40}Y_{i0} + \eta_{41}Y_{i1} + \eta_{42}Y_{i2}$	

  - An MCAR process requires:

⇨ $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

⇨ $\eta_{20} = \eta_{30} = \eta_{31} = \eta_{40}Y_{i0} = \eta_{41}Y_{i1} = \eta_{42} = 0$

Proposed strategies to handle missing data

1. **Always try to avoid or prevent missing data**
   - ⇨ Work on follow-up or tracing missing participants
2. **Collect data on reasons for missing**
   - ⇨ Obtain information about the missing data mechanism
3. **Use descriptive techniques and test MCAR vs MAR**
4. **Where possible do a MAR approach**
   - ⇨ Direct modeling of observed and missing data
   - ⇨ Multiple imputation of missing data : Predicting missingness
5. **Non-ignorable mechannisms (MNAR) require advanced modeling completed by sensitivity analyses**

STATISTICAL ANALYSES WITH MISSING DATA

Typology

➡ There are **3 main families of techniques to handle missing data in the analyses:**

- **The RESTRICTION procedures**
  - These techniques consist in restricting the initial BDD to subjects with complete information
  - **Dangerous advantages: Implemented in too many softwares**
- **The (RE-) WEIGHTING procedures**
  - These techniques compensate the non-response by changing the sampling weights of each subject of the study
- **The IMPUTATION procedures**
  - These many techniques propose plausible values to impute all the missing information contained in a given dataset

Complete Case Analysis [CC]

PRINCIPLE

- *Analyze the complete part of the dataset only, i.e remove all respondents with any missing value from the dataset*

ADVANTAGES

- Simple to compute and easy to explain
- If data is MCAR or MAR, results are unbiased
- P-values, standard error estimates and hypothesis tests are correct

DISADVANTAGES

- Can give bias estimates if data are MNAR
- **Considerable loss of cases** … For a dataset of 10 variables with 10% of missing values (on different subjects), we will only analyze 0.9^{10} complete cases = 34.8%

Available case analysis [Listwise deletion]

PRINCIPLE

- *Use complete cases of EACH variables , or each part of variables to make analyses*

ADVANTAGES

- Easy and applicable for any analysis
- A maximum number of subjects is used / Use all available data

DISADVANTAGES

- Generally valid only under MCAR
- Variable number of subjects used from one sub-analyse to another (difficulty to estimate SE)
- Under MAR estimates maybe seriously biased unless probability of missing data on any independant variables does not depend on values of dependent variable Y

Dummy variable adjustment (in regression analysis)

PRINCIPLE on an example

- *Predict covariate x from y_1*
- *y_1 has 33% missing values (depending on y_0), x is complete*
- *Create dummy variable D: =0 if observed, =1 if missing*
- *Create variable y_1^* = y1 if observed, = c if missing*
- *Regress x on y1* and D: xpred = 1.355 + 0.522y1* + 104.224D*

[Cohen, 1985]

ADVANTAGES

- Easy to implement and very intuitive
- Increase the accuracy of estimates
- Give information about the real mechanism of missingness

DISADVANTAGES

- Give biased estimates and so …
- …The users have to make a choice between reducing bias or increase accuracy

Precisions on weighting procedures

PRINCIPLE

- *Every observed unit is assigned a weight and estimates are based on weighted observations*
- *Weights are derived from probabilities of response*
- *Auxiliary information can be used to make the sample representative for the population*

ADVANTAGES

- Provides sample representative inference
- Can remove some nonresponse bias
- Easy to apply for univariate missing-data patterns and relatively easy for monotone patterns

DISADVANTAGES

- Give biased estimates if data are MNAR
- Very unattractive for arbitrary (multivariate) patterns because each variable needs new estimated weights

LIKELIHOOD BASED METHODS

*"With or without missing data, **the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest** – not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data"*

(Schafer and Graham, 2002)

Objective(s)

- Despite the presence of missing values in classical experimental datasets, **we want to estimate, with the most fiability, unknown population quantity**.

- It implies, for estimates:
  - No or small bias
  - Small SEs (narrow CIs), but close to true value

- Model the distribution of missingness is not a main interest, it must nevertheless be considered in the analyses when necessary (MNAR), in order to minimize its impact on the estimates of interest

Maximum Likelihood Estimation

- **Basic principle:**
Choose as estimates those values that, if true, maximize the probability of observing what has, in fact, been observed *(Allison, 2001)*

- **Likelihood function**
Formula that expresses the probability of the data as function of both the data and the unknown parameter. ML estimates maximize this function

- ML estimates have desirable properties : consistent (approximately unbiased), asymptotically efficient (smallest SEs) and asymptotically normal

- Under the assumption that data are from a multivariate normal distribution, ML can be used to estimate a variety of linear models

What happens when there are missing data ?

- **If the mechanism of missingness is IGNORABLE :**

We obtain the likelihood function by simply using the observed part of the data only:

$X = (X^{obs}, X^{mis})$ $\quad R$: Binary vector of missingness related to X

$$\begin{aligned} V(\theta, \varphi | Y, R, X^{obs}, X^{mis}) &= f_{\theta\varphi}(Y, R | X^{obs}, X^{mis}) \\ &= f_\theta(Y | X^{obs}, X^{mis}) P_\varphi[R | Y, X^{obs}, X^{mis}] \end{aligned}$$

- MCAR or MAR hypothesis implies: $f_\theta(Y|R,X) = f_\theta(Y|X^{obs}).$

$$\begin{aligned} V(\theta, \varphi | Y, R, X^{obs}) &= \int f_{\theta\varphi}(Y, R | X^{obs}, X^{mis}) dX^{mis} \\ &= \int f_\theta(Y | X^{obs}, X^{mis}) P_\varphi[R | Y, X^{obs}] dX^{mis} \\ &= P_\varphi[R | Y, X^{obs}] \int f_\theta(Y | X^{obs}, X^{mis}) dX^{mis} \\ &= P_\varphi[R | Y, X^{obs}] f_\theta(Y | X^{obs}) \end{aligned}$$

The observed-data likelihood

- Gives correct estimates under MAR
- Need to: Write down the function and maximize it
- Finding the observed-data likelihood function is particularly easy for univariate and monotone patterns of missing data because the likelihood function decomposes into separate parts which can be maximized separately
- For general (arbitrary) patterns finding the function is less straightforward
- For general (arbitrary) patterns: EM algorithm

Expectation- Maximization algorithm

- From Dempster, Laird and Rubin, 1977
- Different EM algorithms for different applications: Very *general method to obtain ML estimates* when some of the data are missing
- Application to *multivariate normal distribution*: Estimate the parameters of this model, *i.e.*, the means and the covariance matrix (SDs, correlations)
- *Key idea*: Solve difficult incomplete-data estimation problem by iteratively solving an easier complete-data problem
- Fill in the missing data' with a best guess, then re-estimate the parameters, until convergence

EM algorithm for multivariate normal distribution

An iterative process in which the following two steps are repeated until convergence:

- ❑ **Expectation STEP :**
  Find expected value of the missing data given the observed and current parameter values (imputation with regression permutation)

- ❑ **Maximization STEP :**
  Find new parameter values (ML estimation) given the observed and filled-in data

- ❑ Repeat these 2 steps until parameters stop changing

In this situation, likelihood methods do not assume MAR

Selection models

Models in which we first specify a distribution for the complete data and then propose a manner in which the probability of missingness depends on the (observed) data

Pattern-mixture models

Models that classify respondents by their missingness (patterns) and describe the overal data within each missingness group

⇨ *Procedures required to obtain ML estimates are far from trivial…*

THE IMPUTATION PROCEDURES

"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can legitimately handled in this way and situations where standard estimators applied to real and imputed data have substantial bias."

(Dempster & Rubin, 1983)

PRINCIPLE

"Another way to deal with missing data is to impute all missing values before analysis, using single or multiple imputation methods."

Little and Rubin (2002) suggest 2 approaches to generating this distribution

❑ Explicit modeling

the predictive distribution is based on a formal statistical model (multivariate normal ...)

- **mean /mode imputation** – for any continuous variable missing values are imputed using the mean of the observed values …
  … For categorical variables the mode is used
- **conditional mean imputation (regression imputation)** – missing values are replaced by predicted values from a regression model; least squares, logistic and ordinal regressions are used with continuous, binary and ordered categorical predictors, respectively. *(see Buck, 1960)*
- **stochastic regression imputation** – missing values are imputed by predicted values from a regression model plus a residual

Implicit modeling

the focus is on the algorithm, which implies an underlying model (see *Andridge & Little 2010*)

- ***hot deck imputation*** – missing values are imputed using sampling with replacement from the observed data
- ***substitution*** – nonresponding units are replaced with alternative units not selected into the sample
- ***cold deck imputation*** – missing values are filled in by a constant value from an external source
- ***predictive mean matching*** (*Allison, 2002*) – combination of regression imputation and hot deck method

  1. The method starts with regressing the variable to be imputed Y, on a set of predictors for cases with complete data
  2. On the basis of this regression model predicted values are generated for both the missing and non missing cases
  3. For each case with missing data, a set of cases with complete data that have predicted values of Y that are "close" to the predicted values for the case with missing data is found and from this set of cases one is randomly chosen – its Y value is used to impute the missing case

Few advices for successful imputation

- In the imputation model, the outcome is the incomplete variable
- The imputation requires correct knowing of relationships between the incomplete variable and complete covariates (which suppose MAR process)
- The outcome of the "model of analysis" must be imperatively present as covariate in the model of imputation, as well as any covariate which reflect potential sources of bias
- If we have (by chance) "proxy variables" of the incomplete variable, they have to be also include in the imputation model
- If we have some interactions between the incomplete variable and covariates in the model of analysis, the model of imputation must include interactions between these covariates and the outcome
- The inclusion of unnecessary variables in the imputation model can decrease the effectiveness of the estimator of multiple imputation (*Rubin et Schenker, 1991*)

Last Observation Carried Forward [LOCF]

PRINCIPLE

- *Method of imputation for longitudinal study with monotone pattern*
- *For dropout: If a subject drops out after occasion j, replace y_{ij+1}, y_{ij+2} … by y_{ij}*
- *Equivalent to subject-mean imputation for dropout after first occasion*

ADVANTAGES

- Easy to implement and very intuitive

DISADVANTAGES

- Tends to understate differences in estimated time-trends between treatment and control groups (thought to be "conservative")
- Not necessarily "conservative" because standard errors are biased downward as well
- Especially bad for outcomes that have high variation within a subject

The single imputation procedures

ADVANTAGES

- Generally valid under MCAR and MAR assumptions
- Use available data in a complete way
- Can preserve balanced designs necessary for certain statistical procedures (ANOVA, post-hoc comparisons …)

DISADVANTAGES

- Can be computational resource intensive
- If performed incorrectly it may lead to biased results

!!! After performing imputation complete data analysis procedures can be used for estimation and perform hypothesis testing.
However, standard errors of estimates formed by complete data procedures **do NOT take into account the uncertainty involved in the imputations !!!**

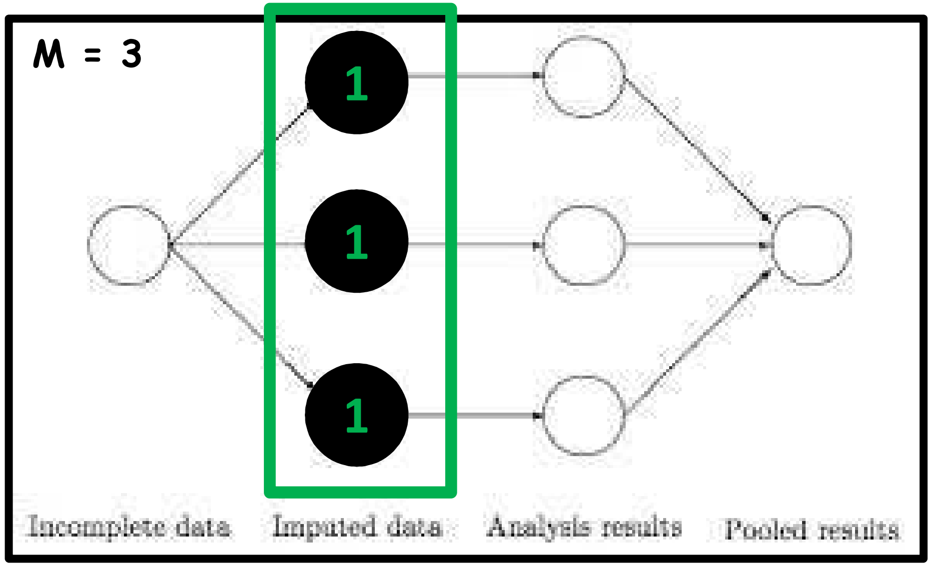
The Multiple Imputation procedures [MI]

- Proposed by D. Rubin in 1978, in response to the single-imputation procedures' main defaults
- Today a standard method of handling missing data (about 1500 citations in PubMed)
- Quantify uncertainty linked to imputed values

The Multiple Imputation process

- There are 3 steps of multiple imputation process (*Yu, 2007*)

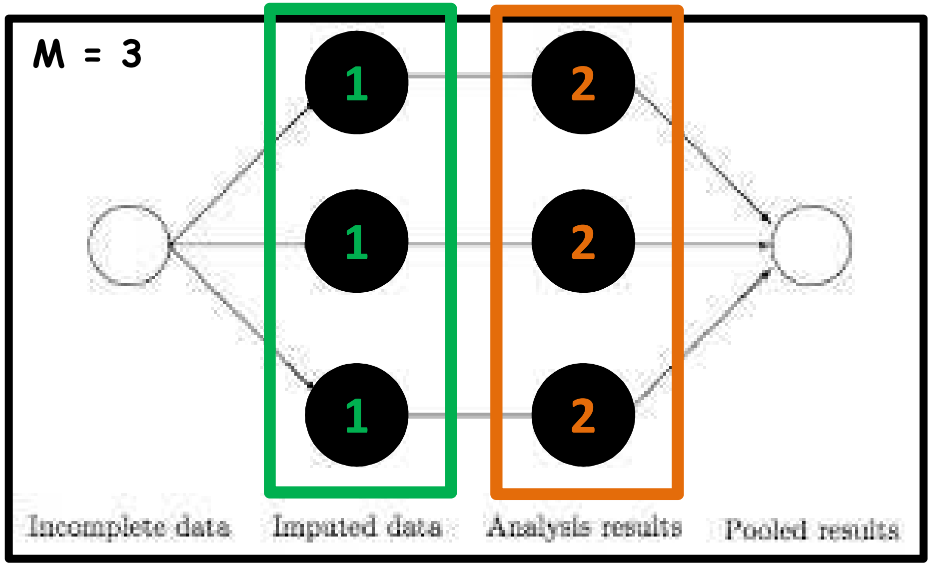
1 Generate M>1 imputed data sets by filling in the missing values with plausible values



The Multiple Imputation process

- There are 3 steps of multiple imputation process (*Yu, 2007*)

1 Generate M>1 imputed data sets by filling in the missing values with plausible values

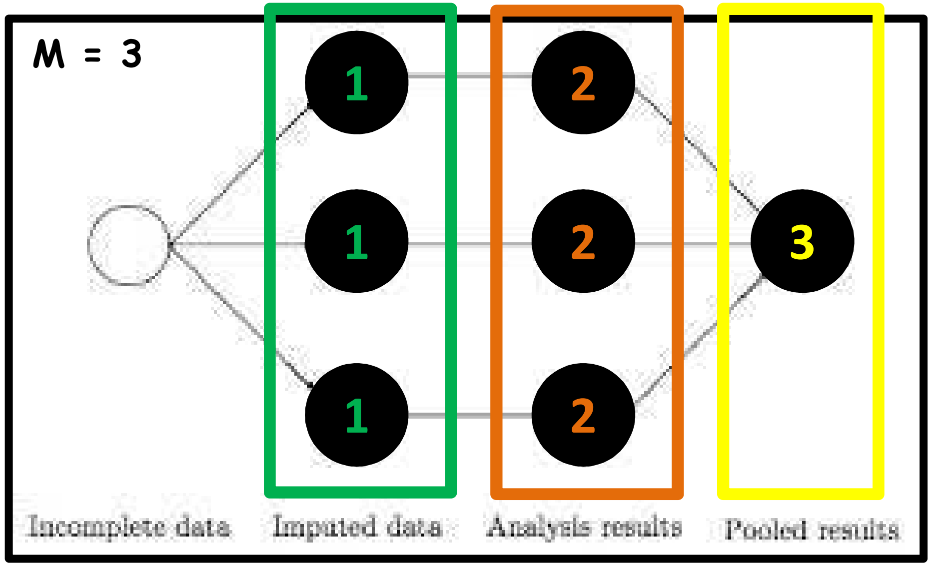


2 Perform standard analyses on each of the M imputed data sets

The Multiple Imputation process

- There are 3 steps of multiple imputation process (*Yu, 2007*)

1 Generate M>1 imputed data sets by filling in the missing values with plausible values



2 Perform standard analyses on each of the M imputed data sets

3 Combine the results from the M analyses

Details of the step 3

- Before this step, suppose that m imputed datasets have been generated
- The same statistical model, with a parameter Q of variance U ,has fitted on each of the M datasets, which generated M estimations of Q

- We want to obtain the most reliable estimate of Q

Notice $\widehat{Q}^{(m)} = \widehat{Q}(Y_{obs}, Y_{miss}^{(m)})$ $U^{(m)} = U(Y_{obs}, Y_{miss}^{(m)})$

⇨ $\overline{Q} = \frac{1}{M}\sum_{m=1}^{M}\widehat{Q}^{(m)}$ *The « pooled » estimate of* Q

⇨ $\overline{U} = \frac{1}{M}\sum_{m=1}^{M}U^{(m)}$ *The within-variance estimate*

⇨ $B = \frac{1}{M-1}\sum_{t=1}^{M}(\widehat{Q}^{(m)} - \overline{Q})^2$ *The between-variance estimate*

Details of the step 3

⇨ $T = \overline{U} + \left(1 + \frac{1}{M}\right) B = \overline{U} + B + \frac{1}{M} B$ *The total variance estimate*

The approximation for inferences could be written : $\frac{Q - \overline{Q}}{\sqrt{T}} \sim t_v$.

Where t_v corresponds to a student distribution with ddl v

⇨ $$v = (M - 1)\left(1 + \frac{\overline{U}}{(1 + M^{-1})B}\right)^2$$

And 2 supplementary definitions:

⇨ $\widehat{r} = \frac{(1 + M^{-1})B}{\overline{U}}$ relative increase in variance due to missingness

⇨ $\widehat{\lambda} = \frac{\overline{U}^{-1} - \frac{v+1}{v+3} T^{-1}}{\overline{U}^{-1}}$ Part of missing information related to Q

How many imputations for reliable analyses ?

- λ is the part of missing information which quantifies the relative information related to a parameter contained in a distribution

- The following table gives the relative efficiency with the formula previously described, in accordance with a given number M of imputation and λ

	λ					
M	0,1	0,3	0,4	0,5	0,7	0,9
3	98	95	94	93	90	88
5	99	97	96	95	94	92
10	100	99	98	98	97	96
20	100	99	99	99	98	98
∞	100	100	100	100	100	100

⇨ For a little number of imputations (from 5 to 10), a high relative efficiency was easily reached

Details of the step 1

- According to van Buuren and Groothuis-Oudshoorn (2010) there are 2 general approaches to simulate multiple imputed values

  - ***joint modeling* (JM)** proposed by Schafer (1997), called **DATA AUGMENTATION**

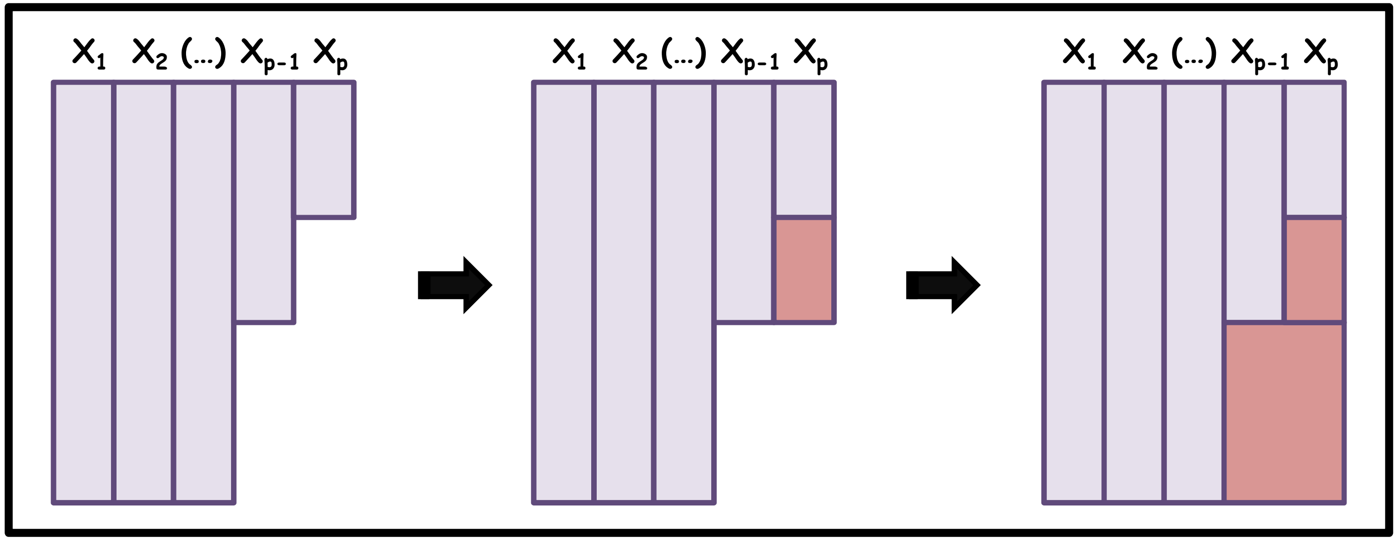
Joint modeling entails specifying a multivariate distribution for the missing data and drawing imputation from their conditional distributions by Markov Chain Monte Carlo (MCMC) techniques (see Schafer 1997, for more details)

  - ***fully conditional specification* (FCS)** developed by van Buuren (2007)

FCS is based on the iterative process that involves specifying a conditional distribution for each incomplete variable.
It does not explicitly assume a particular multivariate distribution, but assumes that one exists and draws can be generated from it using Gibbs sampling (see Yu et al. 2007).

Strategies for multiple imputations

- In a 1st time, MI was reserved to missing data in monotone pattern (dropout, attrition in longitudinal studies)
  - It is so possible to impute missing data by profile, starting with the more complete profiles and ending with the less complete one (see the following figure)
  - At each step, the imputed data was used to impute the remaining missing data



Sequential completion of missing data in a monotone pattern

Strategies for multiple imputations

- In a 2nd time, MI has been extended to non-monotone pattern by using Gibbs sampling approach
- Multivariate Imputation By Chained Equations (MICE) by *van Buuren and Groothuis-Oudshoorn, 2011*
- MICE is the most frequently FCS approach

PRINCIPLE

Starting with an arbitrary 1st imputation, the principle of MICE consists to impute successively the missing values of each incomplete variables conditionnaly to observed data and data previously imputed

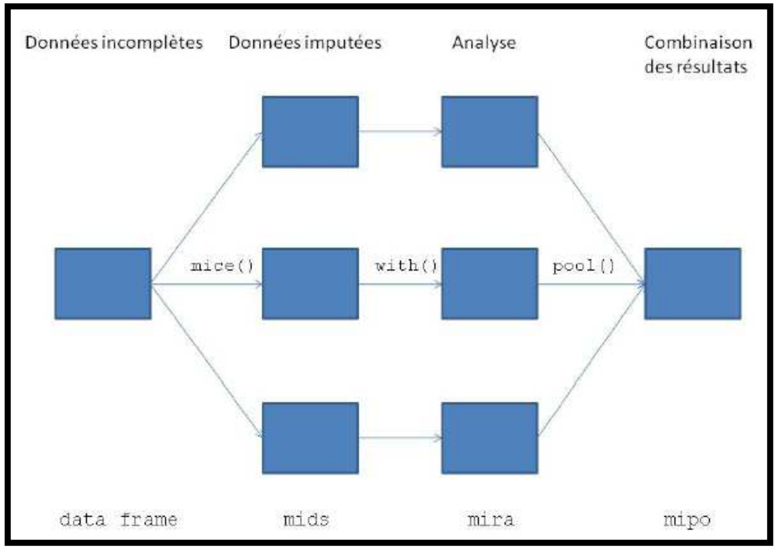
The MICE algorithm

- X_0 is the matrix of complete variables
- $X_1, \ldots, X_p$ the p incomplete variables, and $\theta_1, \ldots, \theta_p$ the p vectors of unknown parameters from the corresponding imputation models

- X_j^{obs} : The observed part of X_j
- X_j^{mis}: The missing part of X_j
- $X_k^{(m)}$: The mth completed vector of X_k, $m = 1, \ldots, M$

$$
\begin{aligned}
\theta_1^{*(t)} &\sim P(\theta_1 | X_0, X_1^{obs}, X_2^{(t-1)}, \ldots, X_p^{(t-1)}) \\
X_1^{*(t)} &\sim P(X_1 | X_0, X_1^{obs}, X_2^{(t-1)}, \ldots, X_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim P(\theta_p | X_0, X_p^{obs}, X_1^{(t)}, \ldots, X_{p-1}^{(t)}) \\
X_p^{*(t)} &\sim P(X_p | X_0, X_p^{obs}, X_1^{(t)}, \ldots, X_{p-1}^{(t)}, \theta_p^{*(t)})
\end{aligned}
$$

Where M is the number of imputations and $X_j^{(t)} = (X_j^{obs}, X_j^{*(t)})$ is the j[th] imputed variables at iteration p



Bilan

ADVANTAGES

- Correct handling of increased uncertainty
- Generating complete datasets that can be analyzed using standard techniques
- Information from data collection process can be used for imputation

DISADVANTAGES

- Imputing missing values with monotone and non-monotone pattern is a lot of work / difficult

Main R packages

Package	Version/ Date	Title	Authors	Description	Basic command
1	2	3	4	5	6
Amelia II	1.2-18 2010-11-04	Amelia II: A Program for Missing Data	James Honaker, Gary King, Matthew Blackwell - Harvard University	Uses a bootstrap+EM algorithm to impute missing values from a dataset and produces multiple output datasets for analysis	amelia(x, m = 5, p2s = 1, frontend = FALSE, idvars = NULL, ts = NULL, cs = NULL, polytime = NULL, splinetime = NULL, intercs = FALSE, lags = NULL, leads = NULL, startvals = 0, tolerance = 0.0001, logs = NULL, sqrts = NULL, lgstc = NULL, noms = NULL, ords = NULL, incheck = TRUE, collect = FALSE, arglist = NULL, empri = NULL, priors = NULL, autopri = 0.05, emburn = c(0,0), bounds = NULL, max.resample = 100, ...)
Hmisc	3.8-3 2010-09-08	Harrell Miscellaneous	Frank E Harrell Jr - Vanderbilt University School of Medicine	Multiple Imputation using Additive Regression, Bootstrapping, and Predictive Mean Matching	aregImpute(formula, data, subset, n.impute=5, group=NULL, nk=3, tlinear=TRUE, ype=c('pmm','regression'), match=c('weighted','closest'), fweighted=0.2, curtail=TRUE, boot.method=c('simple', 'approximate bayesian'), burnin=3, x=FALSE, pr=TRUE, plotTrans=FALSE, tolerance=NULL, B=75)
				Transformations/Imputations using Canonical Variates	transcan(x, method=c("canonical","pc"), categorical=NULL, asis=NULL, nk, imputed=FALSE, n.impute, boot.method=c('approximate bayesian', 'simple'), trantab=FALSE, transformed=FALSE, impcat=c("score", "multinom", "rpart", "tree"), mincut=40, inverse=c('linearInterp','sample'), tolInverse=.05, pr=TRUE, pl=TRUE, allpl=FALSE, show.na=TRUE, imputed.actual=c('none','datadensity','hist','qq','ecdf'), iter.max=50, eps=.1, curtail=TRUE, imp.con=FALSE, shrink=FALSE, init.cat="mode", nres=if(boot.method=='simple')200 else 400, data, subset, na.action, treeinfo=FALSE, rhsImp=c('mean','random'), details.impcat='', ...)

Main R packages

1	2	3	4	5	6
mice	2.4 2010-10-18	Multivariate Imputation by Chained Equations	Stef van Buuren (TNO Quality of Life, Leiden + University of Utrecht) & Karin Groothuis-Oudshoorn (Roessingh RD, Enschede + University Twente)	Multiple Imputation using Fully Conditional Specification	mice(data, m = 5, method = vector("character",length=ncol(data)), predictorMatrix = (1 - diag(1, ncol(data))), visitSequence = (1:ncol(data))[apply(is.na(data),2,any)], post = vector("character", length = ncol(data)), defaultMethod = c("pmm","logreg","polyreg"), maxit = 5, diagnostics = TRUE, printFlag = TRUE, seed = NA, imputationMethod = NULL, defaultImputationMethod = NULL)
mi	0.09-11.03 2010-11-11	Missing Data Imputation and Model Checking	Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima, Maria Grazia Pittau - Columbia University	Multiple Iterative Regression Imputation – the basic command generates a multiply imputed matrix applying the elementary functions iteratively to the variables with missingness in the data randomly imputing each variable and looping through until approximate convergence	mi(object, info, n.imp = 3, n.iter = 30, R.hat = 1.1, max.minutes = 20, rand.imp.method = "bootstrap", run.past.convergence = FALSE, seed = NA, check.coef.convergence = FALSE, add.noise = noise.control())
yaImpute	1.0-12 2010-09-01	yaImpute: An R Package for k-NN Imputation	Nicholas L. Crookston & Andrew O. Finley - Michigan State University	Performs popular nearest neighbor routines for imputation	Find K nearest neighbors: yai(x=NULL, y=NULL, data=NULL, k=1, noTrgs=FALSE, noRefs=FALSE, nVec=NULL, pVal=.05, method="msn", ann=TRUE, mtry=NULL, ntree=500, rfMode="buildClasses") Impute variables from references to targets: impute(object, ancillaryData=NULL, method="closest", method.factor=method, k=NULL,vars=NULL, observed=TRUE,...)

Main R packages

1	2	3	4	5	6
mix	1.0-8 2010-01-03	Estimation/multiple Imputation for Mixed Categorical and Continuous Data	Joseph L. Schafer - The Pennsylvania State University	Imputes Missing Data Under General Location Model	imp.mix(s, theta, x)
norm	1.0-9.2 2010-04-29	Analysis of multivariate normal datasets with missing values	Ported to R by Alvaro A. Novo. Original by Joseph L. Schafer	Imputes missing multivariate normal data	imp.norm(s, theta, x)
cat	0.0-6.2 2009-07-28	Analysis of categorical-variable datasets with missing values	Ported to R by Ted Harding and Fernando Tusell. Original by Joseph L. Schafer	Imputes missing categorical data -performs single random imputation of missing values in a categorical dataset under a user-supplied value of the underlying cell probabilities	imp.cat(s, theta)
pan	0.2-6 2009-04-19	Multiple imputation for multivariate panel or clustered data	Joseph L. Schafer - The Pennsylvania State University	Imputation of multivariate panel or cluster data using the Gibbs sampler algorithm	pan(y, subj, pred, xcol, zcol, prior, seed, iter=1, start)
monomvn	1.8-3 2010-04-23	Estimation for multivariate normal and Student-t data with monotone missingness	Robert B. Gramacy – University of Chicago	Maximum likelihood estimation of the mean and covariance matrix of multivariate normal (MVN) distributed data with a monotone missingness pattern	monomvn(y, pre = TRUE, method = c("plsr", "pcr", "lasso", "lar", "forward.stagewise", "stepwise", "ridge", "factor"), p = 0.9, ncomp.max = Inf, batch = TRUE, validation = c("CV", "LOO", "Cp"), obs = FALSE, verb = 0, quiet = TRUE)

Main R packages

1	2	3	4	5	6
mvnmle	0.1-8 2009-04-17	ML estimation for multivariate normal data with missing values	Kevin Gross, with help from Douglas Bates, North Carolina State University	Finds the maximum likelihood estimate of the mean vector and variance-covariance matrix for multivariate nomal data with missing values	mlest(data, ...)
mitools	2.0.1 2010-05-07	Tools for multiple imputation of missing data	Thomas Lumley – University of Auckland	Tools to perform analyses and combine results from multiple-imputation datasets	MIcombine(results,variances,call=sys.call(), df.complete=Inf,...)
VIM	1.4.2 2010-10-20		Matthias Templ, Andreas Alfons, Alexander Kowarik - Vienna University of Technology	Package introduces new tools for the visualization of missing values in R, which can be used for exploring the data and the structure of the missing values	A lot of commands for visualization and exploring missing data

Selective references for imputation procedures

Allison P. D. (2002), *Missing data*, Series: Quantitative Applications in the Social Sciences 07-136, SAGE Publications, Thousand Oaks, London, New Delhi.

Ambler G., Omar R. Z., Royston P. (2007), *A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome*, "Statistical Methods in Medical Research" 2007; 16: 277–298.

Crookston N. L., Finley A. O. (2008), *yaImpute: An R Package for kNN Imputation*, "Journal of Statistical Software", January 2008, Volume 23, Issue 10.

Horton N. J., Kleinman K. P. (2007), *Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models*, "The American Statistician" 2007, 6 (1): 79-90.

Kenward M. G., Carpenter J. (2007), *Multiple imputation: current perspectives*, "Statistical Methods in Medical Research" 2007; 16: 199–218.

Little R. J. A., Rubin D. B. (2002), *Statistical Analysis with Missing Data*, Wiley, New Jersey.

Molenberghs G., Kenward M. G (2007), *Missing Data in Clinical Studies*, Wiley, England.

Schafer J. L. (1996), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.

Su Y.-S., Gelman A., Hill J., Yajima M. (2011), *Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box*, "Journal of Statistical Software", in press.

van Buuren S., Groothuis-Oudshoorn K. (2011), *MICE: Multivariate Imputation by Chained Equations in R*, „Journal of Statistical Software", in press.

Wayman J. C. (2003), *Multiple Imputation for Missing Data: What Is It And How Can I Use It?*, http://www.csos.jhu.edu/contact/staff/jwayman_pub/ wayman_multimp_aera2003.pdf.

Yu L.-M., Burton A., Rivero-Arias O. (2007), *Evaluation of software for multiple imputation of semi-continuous data*, "Statistical Methods in Medical Research" 2007; 16: 243–258.

A LAST SIMULATED EXAMPLE

❑ Suppose we are interested in the effect of X_1 on Y . Assume, the true relationship of X_1 and Y is :

$$Y = X_1 + X_2 + X_3 + \epsilon, \epsilon \sim N(0, 3^2)$$

… where X_2 and X_3 are related confounding variables. Suppose X_1, X_2, and X_3 are multivariate normally distributed

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & .4 & .3 \\ .4 & 1 & .4 \\ .3 & .4 & 1 \end{pmatrix}\right]$$

- We make roughly 27, 25, 20 percent of the X_1's, X_2' s and X_3's missing respectively with greater probability of lower values being missing. We then estimate the bias, and standard error for our estimate of $\beta 1(=1)$ using several methods of handling the missing data:
  - complete case analysis,
  - mean imputation,
  - EM algorithm to impute data,
  - MI by Bayesian Bootstrap
  - MI by MCMC

- Results

method	bias	sem
all data	0	0.50
complete case analysis	0	0.84
mean imputation	0.08	0.58
EM algorithm	0	0.80
[redacted]		
MI by Bayesian Bootstrap	0.04	0.51
MI by MCMC	0.06	0.72

CONCLUSION

$$\underline{f(\mathbf{Y}_i, \mathbf{R}_i | X_i, \boldsymbol{\theta}, \boldsymbol{\psi})}$$

Selection models: $f(\mathbf{Y}_i | X_i, \boldsymbol{\theta}) \boxed{f(\mathbf{R}_i | X_i, \mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})}$

$\boxed{\text{MCAR}}$	$\rightarrow$	$\boxed{\text{MAR}}$	$\rightarrow$	$\boxed{\text{MNAR}}$
$f(\mathbf{R}_i \mid X_i, \boldsymbol{\psi})$		$f(\mathbf{R}_i \mid X_i, \mathbf{Y}_i^o, \boldsymbol{\psi})$		$f(\mathbf{R}_i \mid X_i, \mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})$
CC?		**direct likelihood!**		joint model!?
LOCF?		EM		sensitivity analysis?!
single imputation?		**MI!**		
⋮		**WGEE!**		

Pattern-mixture models: $f(\mathbf{Y}_i | X_i, \mathbf{R}_i, \boldsymbol{\theta}) f(\mathbf{R}_i | X_i, \boldsymbol{\psi})$

Shared-parameter models: $f(\mathbf{Y}_i | X_i, \mathbf{b}_i, \boldsymbol{\theta}) f(\mathbf{R}_i | X_i, \mathbf{b}_i, \boldsymbol{\psi})$

Resources on missing data in general

- Little and Rubin (2002) *Statistical Analysis with Missing Data, Second edition.* (New York: Wiley)
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data* (London: Chapman & Hall)
- Allison, P.D. (2001) *Missing Data* (Thousand Oaks: Sage)
- Schafer, J.L. and Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychological Methods*

Resources on missing data in longitudinal studies

- Little, R.J. (1995) Modeling the dropout mechanism in repeated-measures studies. *JASA*
- Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data* (New York: Springer).