Données Manquantes

A. Latouche

Plan

- Contexte
- Classification des mécanismes de manquement
- Méthodes Biaisées
- Méthodes d' imputation
- Imputation et Données longitudinales

Problématique Générale

L'ensemble des données avec lequel on doit travailler n'est pas toujours complet (euphémisme)

	Variables				
	1	2	3	...	P
1		NA			
2					
3			NA		
.					
N					NA

- Données manquantes
  - Variable à expliquer
  - Variable(s) explicative(s)
- Différents cadres sont possibles
  - Essais thérapeutique
    - Sortis de l'étude
    - Perdus de vue

    $\rightarrow$ Absence de valeur pour le critère de jugement
  - Enquêtes
    - Non réponse totale
    - Non réponse partielle

Impacts

- Perte d'information non pertinente et/ou non informative : Impact nul
- Perte d'information pertinente et/ou informative
  - Impact fonction du taux de NA
  - Biais possible dans l'estimation de la précision et de l'exactitude
- Solutions
  - Cas complet : Lorsque la proportion de NA de l'échantillon est **faible**
  - Utiliser une méthode adaptée

Impacts

Analyse univariée

	Variables				
	1	2	3	4	5
1		NA		NA	
2					
3	NA				
4		NA		NA	NA
5					
6			NA	NA	
% NA					

NA : observations exclues de l'analyse

Impacts

Analyse multivariée

	Variables				
	1	2	3	4	5
1		NA		NA	
2					
3	NA				
4		NA		NA	NA
5					
6			NA	NA	

33,3 % d'observations complètes

Classification des Données Manquantes

MCAR : manquant complètement au hasard

- La probabilité qu'une observation soit incomplète est constante
- le fait de ne pas avoir la valeur pour une variable X_i est indépendant des autres variables $X_{j \neq i}$

Exemple

- L'appareil mesurant la variable n'a plus de batterie
- Le tube contenant le prélèvement est tombé : pas d'observation

Au final, très rare en pratique courante

Classification des Données Manquantes

MAR : manquant au hasard

- La probabilité qu'une observation soit incomplète ne dépend que de valeurs observées (pas de valeurs manquantes)
- i.e le fait de ne pas avoir la valeur pour une variable X_i est dépendant d'une autre ou d'autres variables $X_{j \neq i}$ observées

Exemple

- Elle n'est pas la même pour tous les sujets (mais ne dépend que de l'information observée)

Classification des Données Manquantes

NMAR : ne manquant pas au hasard (informative)

- La probabilité qu'une observation soit incomplète dépend de valeurs non observées
- Elle n'est pas aléatoire i.e le fait de ne pas avoir la valeur pour une variable X_i observée est dépendant d'une autre ou d'autres valeurs non observées des variables $X_{j \neq i}$ observées

Exemple

- Elle n'est pas la même pour tous les sujets
- La proportion de données manquantes dépends du temps de suivi (cohorte prospective)

Exemple :

- Personnes avec un revenu important refusent de le dévoiler
- Un patient meurt après être écrasé par un bus. Les informations postérieures à l'accident sont MCAR, si ce patient participait à un essai de psychiatrie cela peut indiquer une mauvaise réponse au traitement (MNAR).

Méthodes de Traitement

- Analyse de données complètes
- Imputation simple
- Imputation multiple

Analyse de données complètes

Au lieu de travailler sur

	Variables				
	1	2	3	4	5
1		NA		NA	
2					
3	NA				
4		NA		NA	NA
5					
6			NA	NA	

Analyse de données complètes

On travaille sur les cas complets

	Variables				
	1	2	3	4	5
2					
5					

Analyse de Données Complètes

- Stratégie la plus courante
- Généralement imposée par les logiciels
- Proportion d'observations complètes peut être faible même si, pour chaque variable, la probabilité qu'une donnée soit observée est grande
- Résultats non biaisés si les données sont MCAR

Sinon **biais importants**

Imputation Simple

- Au lieu de travailler sur les données complètes
- On remplace (impute) chaque NA par une donnée prédite ou simulée
  - Prédiction grâce à un modèle de régression
  - Simulation issue de la loi des données observée

Imputation Simple

- Hypothèse d'un processus d'observation MAR
- Produit une valeur artificielle pour remplacer la valeur NA
- Les informations disponibles sur les individus qui ne fournissent qu'une réponse partielle peuvent être utilisées comme variables auxiliaires pour améliorer la qualité des valeurs imputées

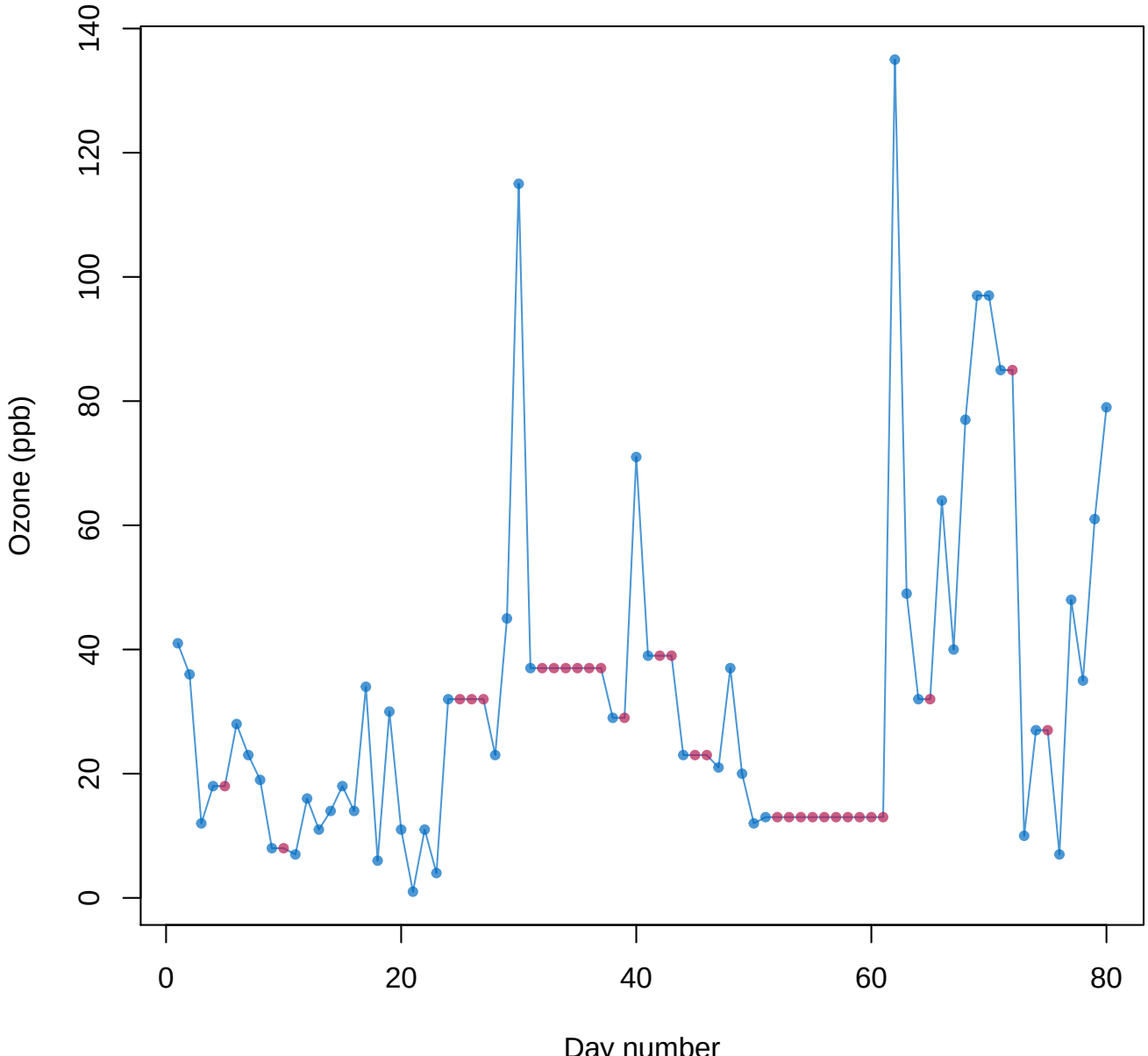
Imputation Simple : LOCF

- Last Observation Carried Forward
- Lors de mesures répétées
- Suppose que la vraie valeur reste inchangée depuis la dernière mesure
- Si pas de mesure disponible pendant le suivi, la valeur initiale est utilisée

id	$X(t_0)$	$X(t_1)$	$X(t_2)$	$X(t_3)$	$X(t_4)$
1	18	20	NA	15	20
2	24	NA	NA	NA	17

Biaisé

LOCF

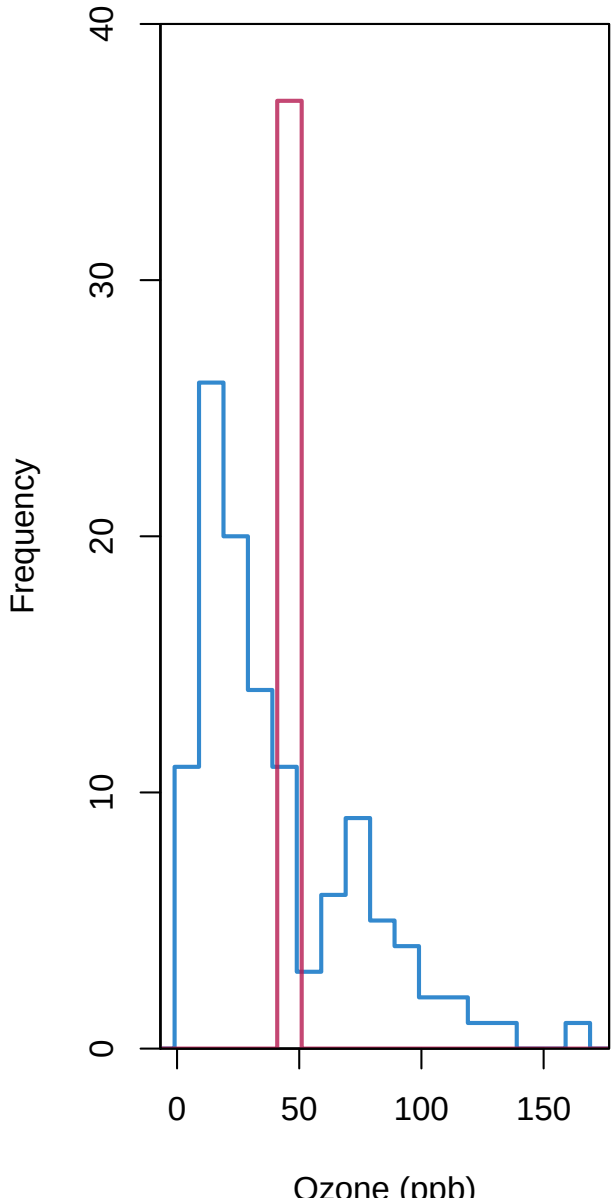


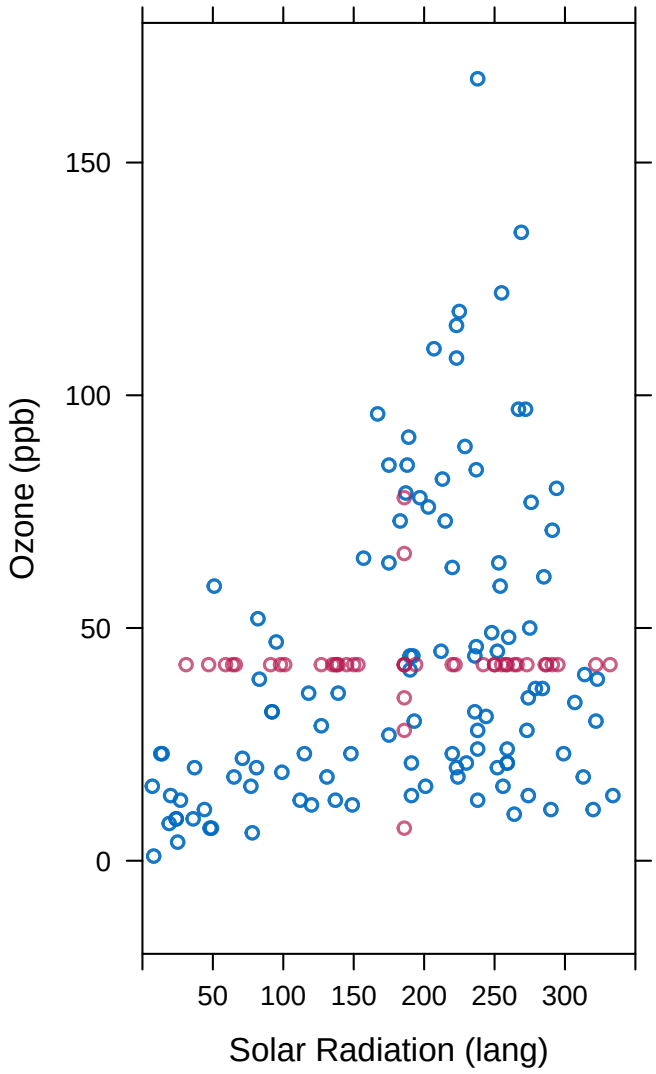
Imputation Simple : par la Moyenne

- Remplacer une valeur manquante par la moyenne des mesures disponibles
- La même pour toutes les NA d'une même variable
- Estimations non biaisées si les données sont MCAR

Sous estimation de la variance

Imputation par la moyenne





IS : par un modèle de Régression

- Remplacement d'une valeur manquante Y_i par une valeur prédite $Y*$ obtenue par régression de Y sur X_1, X_2 ...
- Possibilité d'ajouter un aléa à la prédiction
- Estimation ponctuelle correcte
- Variance sous-estimée

IS : par un modèle de Régression

Exemple : régression linéaire simple

X	Y
	NA
	NA
	NA
	NA

- Modélisation sur les cas complets $i = 1, 2, 3$

$$Y_i = \hat{a} + \hat{b}X_i$$

- Imputation par la prédiction du modèle de régression $i = 4, 5, 6, 7$

$$Y_i^\star = \hat{a} + \hat{b}X_i$$

Imputation Multiple

- Méthode consistant à créer plusieurs valeurs possibles d'une valeur manquante

- Les buts sont
  - De refléter correctement l'incertitude des NA
  - De préserver les aspects importants des distributions
  - De préserver les relations importantes entre les variables

- Les buts ne sont pas
  - De prédire les données manquantes avec la plus grande précision
  - De décrire les données de la meilleur façon possible

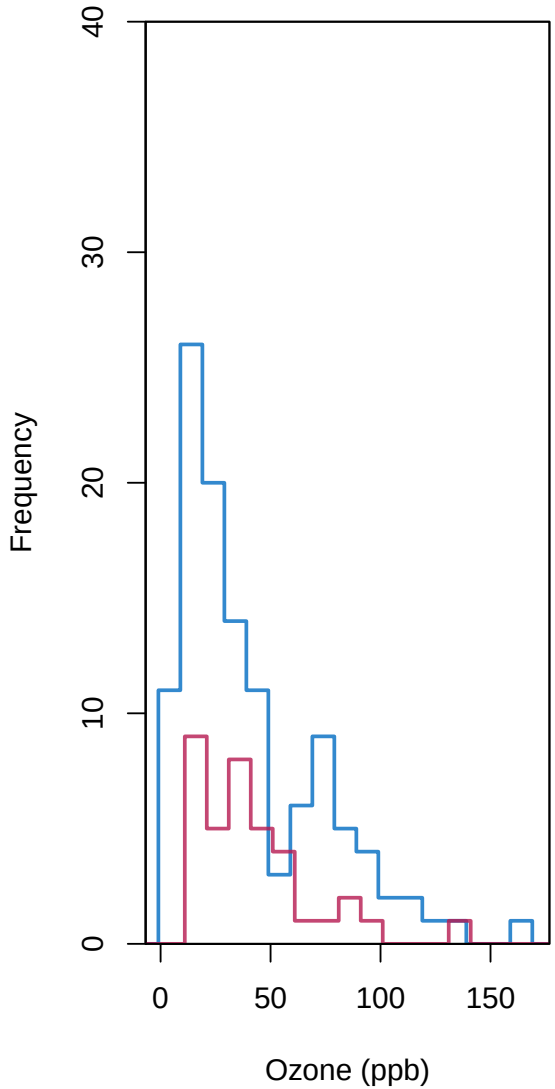
MI

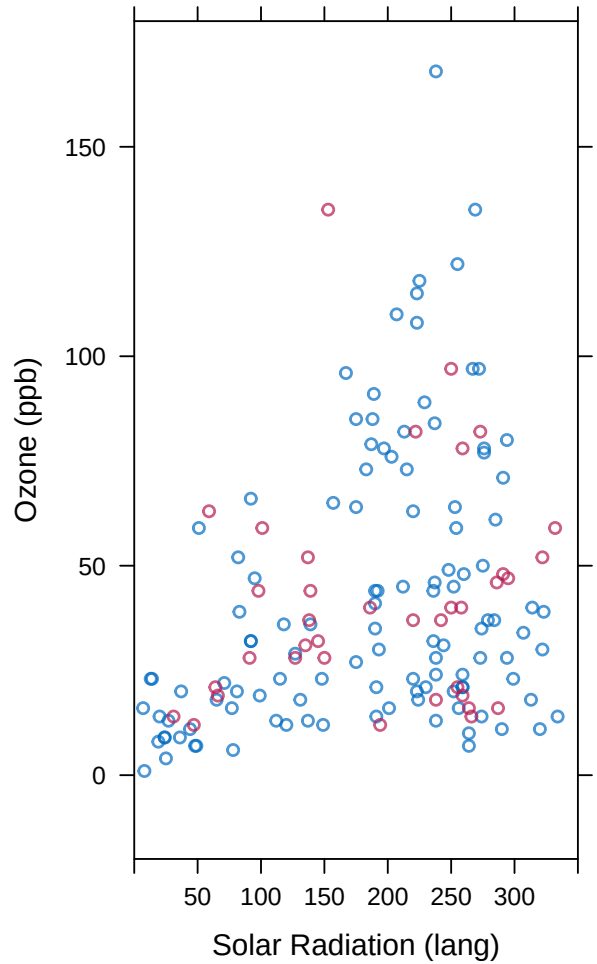
- Remplacer les valeurs manquantes par M valeurs plausibles.

- Combiner les résultats des analyses sur chaque jeu de données.

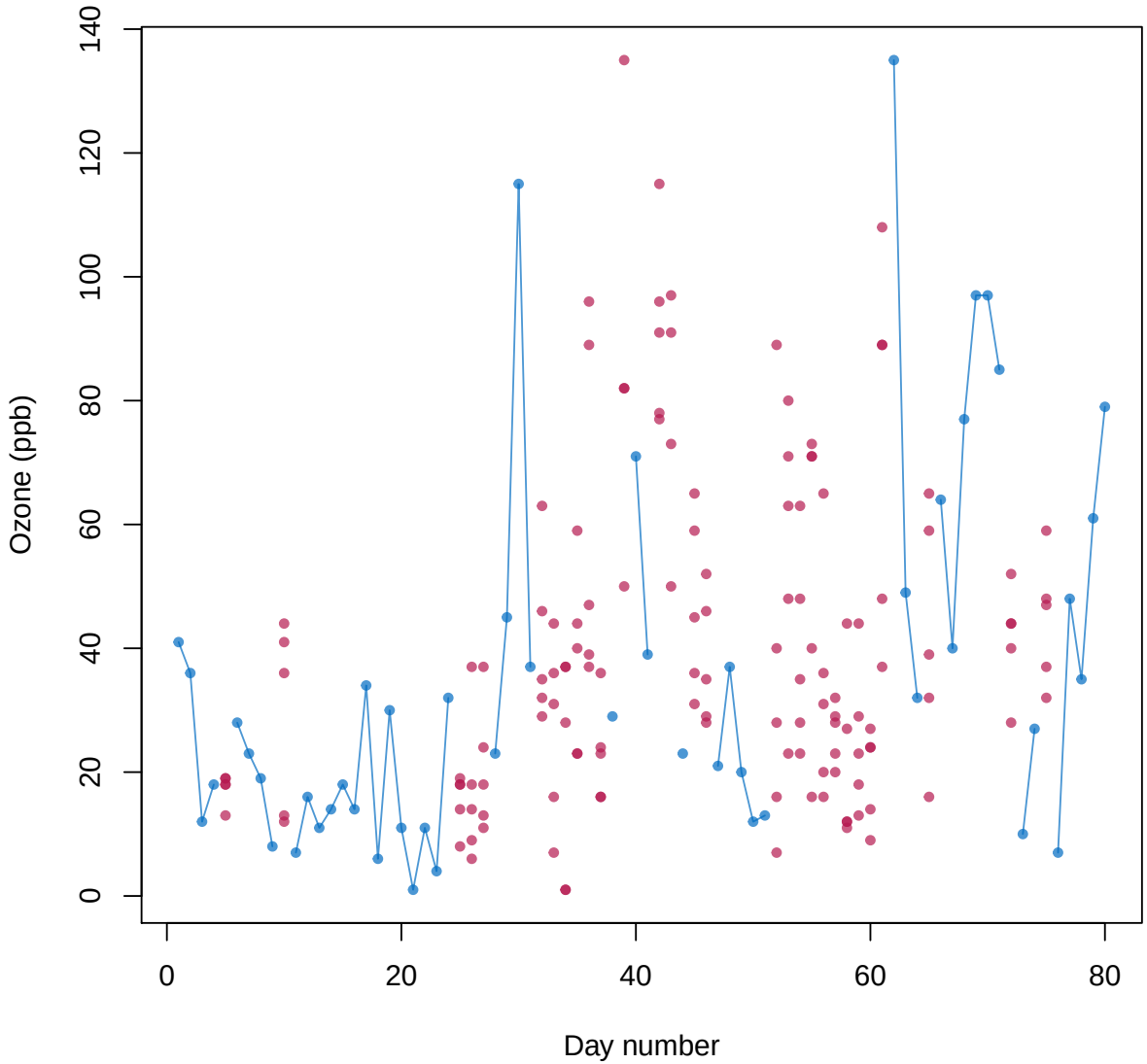
Sous MAR et un modèle d'imputation correct : correction du biais et gain de précision.

MI : Ozone


Frequency
0
10
20
30
40
0
50
100
150
Ozone (ppb)


Ozone (ppb)
0
50
100
150
50
100
150
200
250
300
Solar Radiation (lang)

MI : Ozone

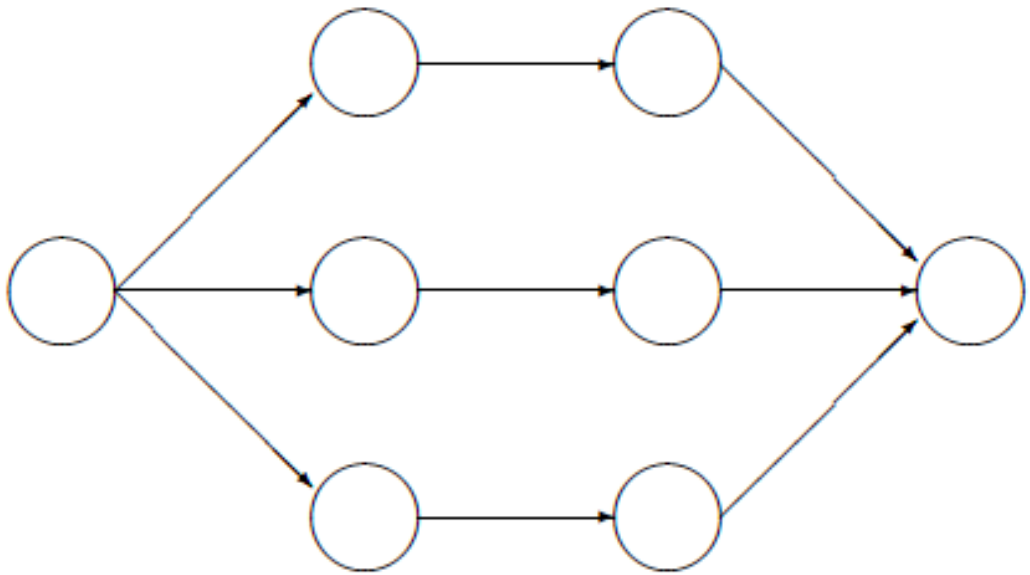


Mise en oeuvre

National Health and Nutrition Examination Survey

- 25 sujets, 4 variables
- age : groupes d'âge : 1=20-39, 2=40-59, 3=60+
- bmi : index de masse corporelle (kg/m2)
- hyp : hypertension : 1=non, 2=oui
- chl : cholestérol total (mg/dL)

Imputation Multiple



National Health and Nutrition Examination Survey

	age	bmi	hyp	chl
1	1.00			
2	2.00	22.70	1.00	187.00
3	1.00		1.00	187.00
4	3.00			
5	1.00	20.40	1.00	113.00
6	3.00			184.00

Structure des données manquantes

Codage : 1 observé, 0 manquant

	age	hyp	bmi	chl	
13	1	1	1	1	0
1	1	1	0	1	1
3	1	1	1	0	1
1	1	0	0	1	2
7	1	0	0	0	3
	0	8	9	10	27

Ex: 13 lignes sans données manquantes, 3 lignes où chl est manquante, 27 NA au total

Valeurs imputées bmi

Obs NA	Valeur Imputée				
	1	2	3	4	5
1	33.20	22.70	27.40	27.50	30.10
3	30.10	22.00	30.10	22.00	28.70
4	21.70	20.40	27.50	30.10	27.20
6	24.90	21.70	25.50	30.10	24.90
10	25.50	27.50	22.50	27.20	22.70
11	35.30	35.30	29.60	22.00	26.30
12	35.30	25.50	27.50	33.20	25.50
16	35.30	35.30	27.40	27.40	29.60
21	27.20	30.10	27.50	22.50	27.20

Régression Linéaire

- Variable réponse Y=chl

- Variable explicatives X_1=âge et X_2=hyp

Régression Linéaire

Estimation

$$\text{chl} = 136.7 + 22.32 \times \text{age} + 15.98 \times \text{hyp}$$

Fraction de l'information due à la non-réponse

Intercept	age	hyp
0.3505934	0.3713103	0.6027942

Nombre de NA

age	bmi	hyp	chl
0	9	8	10

Regression Linéaire : Cas complet

	age	bmi	hyp	chl
2	2.00	22.70	1.00	187.00
5	1.00	20.40	1.00	113.00
7	1.00	22.50	1.00	118.00
8	1.00	30.10	1.00	187.00
9	2.00	22.00	1.00	238.00
13	3.00	21.70	1.00	206.00
14	2.00	28.70	2.00	204.00
17	3.00	27.20	2.00	284.00
18	2.00	26.30	2.00	199.00
19	1.00	35.30	1.00	218.00
22	1.00	33.20	1.00	229.00
23	1.00	27.50	1.00	131.00
25	2.00	27.40	1.00	186.00

Régression Linéaire : Cas complet

$$\text{chl} = 111.47 + 33.19 \times \text{age} + 20.04 \times \text{hyp}$$

vs.

$$\text{chl} = 136.7 + 22.32 \times \text{age} + 15.98 \times \text{hyp}$$

Erreurs dans les prédictions

Essai clinique d'un traitement pour l'insomnie

- 962 individus randomisés traitement ou témoin.
- Réponses au questionnaire au cours de 5 périodes:

  - Temps de réveil pendant la nuit (WASO)
  - Temps total de sommeil (TST)
  - Délai d'endormissement (SOL)

- Critére clinique : qualité du sommeil pendant la cinquiéme période.

Données manquantes au niveau quotidien

Table : Moyenne (é.t.) du nombre effectif de mesures disponibles pour chaque individu par visite et par score pour le groupe témoin.

	Baseline	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
WASO	6.8 (1.2)	12.4 (2.4)	12.0 (2.7)	11.9 (2.9)	17.2 (4.8)	16.7 (4.8)
TST	6.8 (1.2)	12.5 (2.3)	12.1 (2.6)	12.0 (2.9)	17.3 (4.9)	16.9 (4.9)
SOL	6.8 (1.1)	12.5 (2.3)	12.1 (2.7)	12.0 (2.9)	17.4 (4.8)	16.9 (4.7)

Table : Moyenne (é.t.) du nombre effectif de mesures disponibles pour chaque individu par visite et par score pour le groupe de traitement.

	Baseline	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
WASO	6.8 (1.3)	12.2 (2.7)	11.9 (3.0)	11.8 (2.9)	16.9 (5.0)	16.7 (4.7)
TST	6.8 (1.3)	12.3 (2.6)	12.0 (2.9)	11.8 (2.9)	17.1 (4.9)	16.7 (4.7)
SOL	6.8 (1.3)	12.3 (2.6)	12.1 (2.9)	11.9 (2.7)	17.2 (4.8)	16.9 (4.6)

Données manquantes au niveau période: les sorties d'étude étaient la cause principale

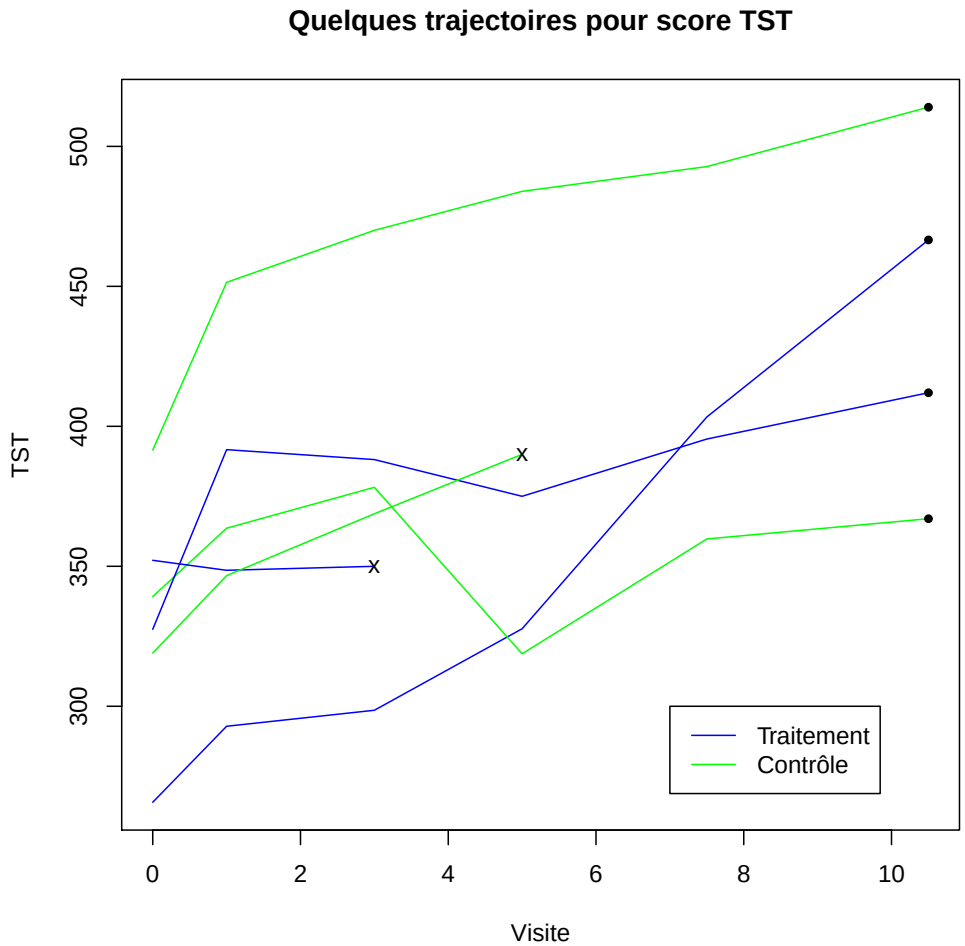
Table : Pourcentage de données manquantes par visite et par score pour le groupe témoin.

	Baseline	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
WASO	0.0	0.6	7.2	14.5	19.4	24.1
TST	0.0	0.6	8.1	15.4	19.7	24.6
SOL	0.0	0.6	7.5	14.8	19.7	24.6

Table : Pourcentage de données manquantes par visite et par score pour le groupe de traitement.

	Baseline	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
WASO	0.0	0.3	6.2	13.0	16.7	21.4
TST	0.0	0.3	6.2	12.2	17.0	21.2
SOL	0.0	0.3	6.0	12.6	16.7	20.9

Modélisation en présence de données incomplètes



L'exclusion des observations incomplètes dans l'analyse peut entrainer des biais de sélection ainsi qu'une perte de précision dans les estimations.

La typologie des données manquantes

Pour chaque individu $i \in \{1, \ldots, n\}$, soit

- $Y_i = (Y_{i1}, \ldots, Y_{iJ})$ le vecteur des J réponses de l'individu i,
- $R_{ij} = 1$ si Y_{ij} est observée, $R_{ij} = 0$ dans le cas contraire,
- Y_i^o la partie observée de Y_i et Y_i^m la partie non-observée
- X_i un vecteur de covariables complètement observé.

- *Données manquantes complètement au hasard* (MCAR):

$$P(R_{ij} = 0|X_i, Y_i^o, Y_i^m) = P(R_{ij} = 0) = \alpha.$$

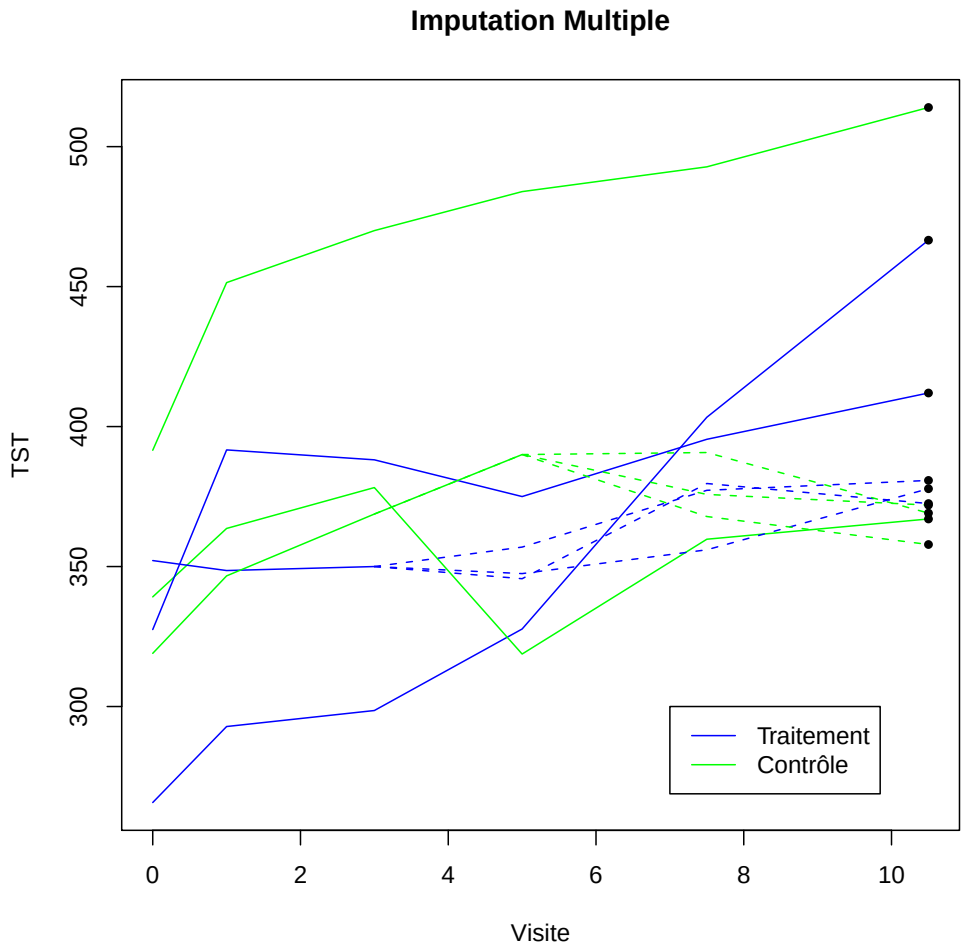
- *Données manquantes au hasard* (MAR):

$$P(R_{ij} = 0|X_i, Y_i^o, Y_i^m) = P(R_{ij} = 0|X_i, Y_i^o) = \alpha(X_i, Y_i^o).$$

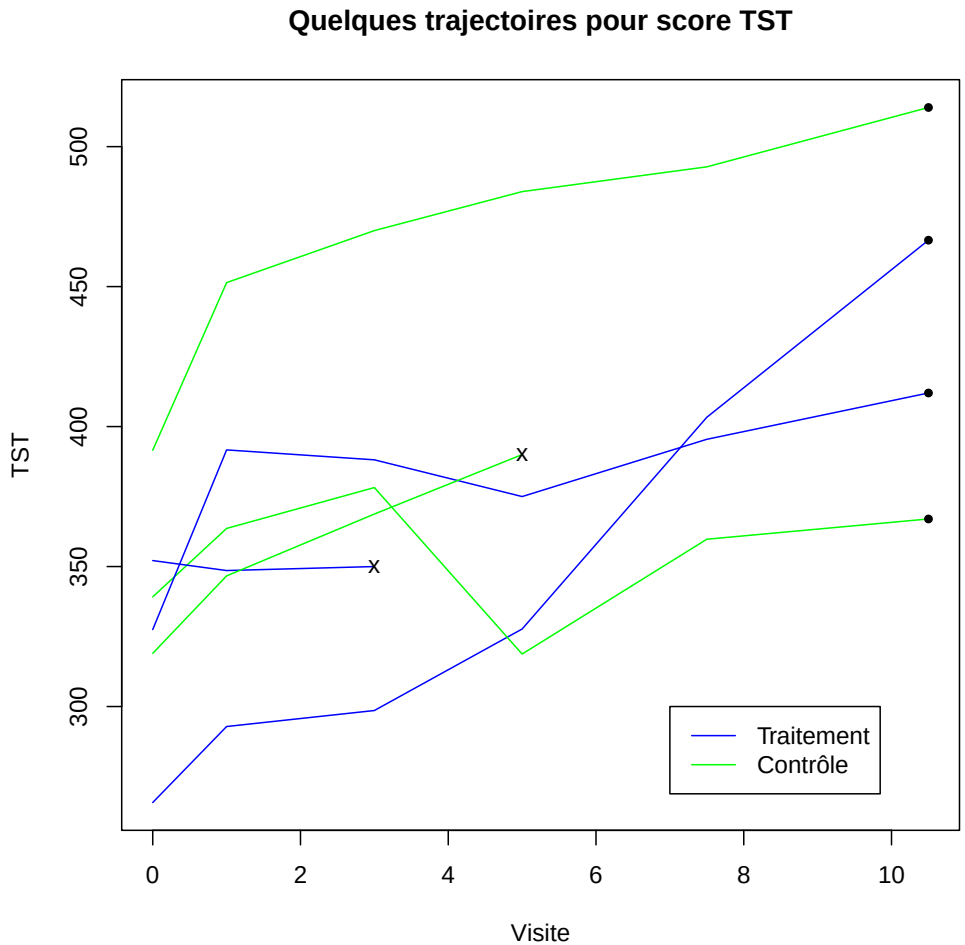
- *Données manquantes non au hasard* (MNAR):

$$P(R_{ij} = 0|X_i, Y_i^o, Y_i^m) = \alpha(X_i, Y_i^o, Y_i^m).$$

Approche proposée : l'imputation multiple



Approche proposée : l'imputation multiple



Estimateur de l'imputation multiple

Estimer le paramètre θ d'intérêt et sa variance sur chaque jeu de données : $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}$ et $\hat{V}_\theta{}^{(1)}, \ldots, \hat{V}_\theta{}^{(M)}$.

L'estimateur de l'imputation multiple de θ est donné par:

$$\hat{\theta}^* = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)}.$$

Estimateur de l'imputation multiple

Sa variance est la somme de la variance intra-imputation et la variance inter-imputation des $\hat{\theta}^{(m)}$'s:

$$V_\theta^* = W + B.$$

On l'estime par

$$\hat{V}_\theta^* = \hat{W} + \left(1 + \frac{1}{M}\right) \hat{B}$$

o

$$\hat{W} = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_\theta{}^{(m)}$$

et

$$\hat{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \hat{\theta}^*)(\hat{\theta}^{(m)} - \hat{\theta}^*)'.$$

MICE

Most of the time, missing values occur in several variables.

- A flexible approach in this case is *multiple imputation by chained equations* (MICE), also known as *fully conditional specification*.
- Let the data be represented by the $n \times p$ matrix Y,
- let Y_j be the j-th column of Y
- The observed data are collectively denoted by Y^{obs} and Y^{mis} collectively denotes the missing data.

The MICE algorithm

Initially, all missing values are filled in by random draws with replacement taken from the observed data.

1. The first variable Y_1 is regressed on all other variables Y_1, restricted to Y_1^{obs}.
2. Values for Y_1^{mis} are drawn from the posterior predictive distribution of Y_1.
3. The next variable Y_2 is regressed on all other variables Y_2, using the imputed values of Y_1.
4. The regression model is again restricted to observed Y_2. Y_2^{mis} is filled by draws from the posterior predictive distribution of Y_2.
5. This process is repeated for all other variables with missing values $3, \ldots, p$ in turn.

This cycle is usually repeated several times (e.g., 10 to 20 times) in order to stabilise the results. This results in an imputed data set. The whole process is repeated m times to give m imputed data sets.