Traitement des Données Manquantes

Pr Roch Giorgi

roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

Problématique Générale (1)

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

Variables

Observations

	1	2	3	…	P
1		NA			
2					
3			NA		
.					
.					
.			NA		
N					NA

Problématique Générale (2)

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

    } Absence de valeur pour le critère de jugement
  - ✓ Enquêtes
    - Non réponse totale
    - Non réponse partielle

Problématique Générale (3)

- 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
  - ✓ Ne rien faire
    - Lorsque la proportion de NA de l'échantillon est faible ($\leq$ 5 %)
  - ✓ Utiliser une procédure adaptée de remplacement des NA

Impacts (1)

- Analyse univariée

Variables

Observations

	1	2	3	4	5
1		NA		NA	
2					
3	NA				
4		NA		NA	NA
5					
6			NA	NA	
% NA	16,7	33,3	16,7	50,0	16,7

NA observations exclues de l'analyse

Impacts (2)

- Analyse multivariée

Variables

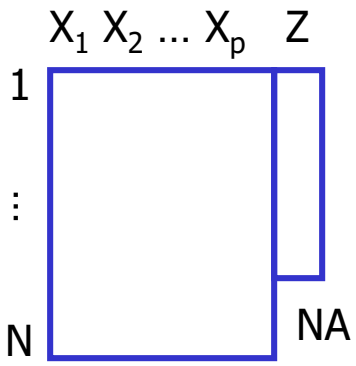
Observations

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

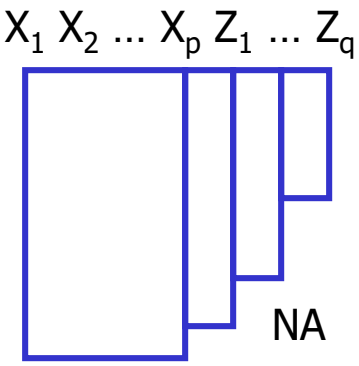
33,3 % d'observations complètes

NA observations exclues de l'analyse

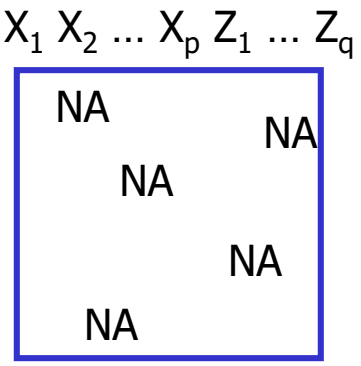
Pattern des Données Manquantes



Univarié



Monotone



Général Multivarié

Classification des Données Manquantes (1)

- MCAR : manquant complètement au hasard
  - ✓ La probabilité qu'une observation soit incomplète est une constante
  - ✓ i.e le fait de ne pas avoir la valeur pour une variable X_i est indépendant des autres variables $X_{j \neq i}$

- Exemple
  - ✓ X_1 = âge ; X_2 = sexe ; X_3 = glycémie
  - ✓ La probabilité que l'âge soit NA ne dépend ni du sexe, ni des valeurs de glycémie
  - ✓ Elle est la même pour tous les sujets

Classification des Données Manquantes (2)

- 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
  - ✓ X_1 = âge ; X_2 = sexe ; X_3 = glycémie
  - ✓ La probabilité que l'âge soit NA ne dépend que du sexe et des valeurs de la glycémie (valeurs observées)
  - ✓ Elle n'est pas la même pour tous les sujets

Classification des Données Manquantes (3)

- 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
  - ✓ X_1 = âge ; X_2 = sexe ; X_3 = glycémie
  - ✓ La probabilité que l'âge soit NA dépend des valeurs manquantes pour le sexe et la glycémie (valeurs non observées)
  - ✓ Elle n'est pas la même pour tous les sujets

Commentaires / Remarques (1)

- MAR et NMAR
  - ✓ MAR et méthode d'analyse pertinente ⇒ inférences correctes (seules les informations observées sont utiles)
- MCAR et MAR
  - ✓ On parle de processus ignorable ou non-informatif
  - ✓ Si analyse correcte, ne nécessite pas de modéliser le processus d'observation
- NMAR
  - ✓ On parle de processus non-ignorable ou informatif
  - ✓ Inférence sur la population étudiée nécessite
    - Poser des hypothèses fortes
    - Ou d'obtenir des informations complémentaires
  - ✓ Nécessité de modéliser le processus d'observation

Commentaires / Remarques (2)

- Processus MAR ou NMAR ?
  - ✓ Cela revient à se demander si la probabilité qu'une observation soit complète dépend ou non d'une quantité inconnue $\Rightarrow$ impossible
  - ✓ Cas particulier
    - Données MAR lorsqu'il est prévu un recueil incomplet dans une situation définie
    - Exemple

      X_1, X_2, …, X_k recueillies pour tous les sujets

      X_{k+1}, X_{k+2}, …, X_p recueillies uniquement chez certains sujets en dépendant uniquement de X_1, X_2, …, X_k
  - ✓ Hypothèse H_1 = MCAR vs H_0 = MAR peut être testée
    - Si H_1 = MCAR n'est pas rejetée, l'hypothèse d'un processus ignorable est plus crédible

Stratégies d'Analyse

- Fonction de plusieurs questions
  - ✓ Quelle(s) hypothèse(s) formule-t-on sur le processus d'observation ?
  - ✓ En fonction de cela, quelles données va-t-on modéliser ?
  - ✓ Si l'on n'écarte pas la possibilité de données NMAR, comment tenir compte du risque de biais ?

Méthodes de Traitement

- Analyse de données complètes
- Indicateur de données manquantes
- Analyse pondérée
- Imputation simple
- Imputation multiple

Analyse de Données Complètes (1)

Au lieu de travailler sur …

	1	2	3
1			
2	NA		
3		NA	
4			
5			
6			
7		NA	
8			
9			NA
10			
11			NA
12			
13			
14			

	1	2	3
1			
4			
5			
6			
8			
10			
12			
13			
14			

… l'analyse ne porte que sur les enregistrements complets

Analyse de Données Complètes (2)

- 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
  - ✓ Mais diminution de la précision et de la puissance
- Sinon biais importants

Exemple : Risque prématurité

(Chavance-Manfredi, 2000)

- Risque de prématurité
- Facteurs étudiés
  - ✓ Antécédents obstétricaux pathologique
  - ✓ Consommation de tabac ≥ à 5 cigarettes par jour pendant le troisième trimestre de grossesse (7 % de NA)

	Antécédent obstétrical pathologique	
	Log(ORa)	ET(Ora)
Données complètes	1,46	0,12

ORa : Odds Ratio ajusté sur la consommation de tabac ≥ 5 cig./jour pendant le troisième trimestre de la grossesse

Indicateur de Données Manquantes (1)

Au lieu de travailler sur …

	1
1	1
2	NA
3	2
4	2
5	2
6	1
7	NA
8	2
9	NA
10	1
11	NA
12	1
13	2
14	2

on ajoute une modalité à la variable catégorielle incomplètement observée

	1
1	1
2	**9**
3	2
4	2
5	2
6	1
7	**9**
8	2
9	**9**
10	1
11	**9**
12	1
13	2
14	2

et l'analyse portera sur tous les enregistrements

Indicateur de Données Manquantes (2)

- Suppose des données MCAR ou MAR
- Peut améliorer la précision de certains estimateurs
- Permet d'apprécier le risque de biais
  - ✓ Une interaction significative entre l'indicatrice de données manquantes et une variable explicative signale l'existence d'un problème
- Mais ne protège pas contre le risque de biais

Exemple : Risque prématurité

(Chavance-Manfredi, 2000)

	Antécédent obstétrical pathologique	
	Log(ORa)	ET(Ora)
Données complètes	1,46	0,12
Indicateur de données manquantes	1,36	0,11

Analyse Pondérée (1)

- Pour des données MAR
- Estimation de la probabilité qu'une observation soit complète pour chaque combinaison des variables influençant cette probabilité
- Exemple
  - ✓ Échantillon aléatoire représentatif
  - ✓ 2 catégories de sujets avec taux de réponses 50 % et 100 %
  - ✓ Chacune des observations de la première catégorie représente 2 fois plus de sujets qu'une observation de la deuxième

Analyse Pondérée (2)

- Affectation pour chaque observation complète d'un poids inversement proportionnel à la probabilité d'effectuer une observation complète
- Correction du biais éventuel
- Augmentation de l'imprécision
- Nécessite une correction particulière pour estimer la variance

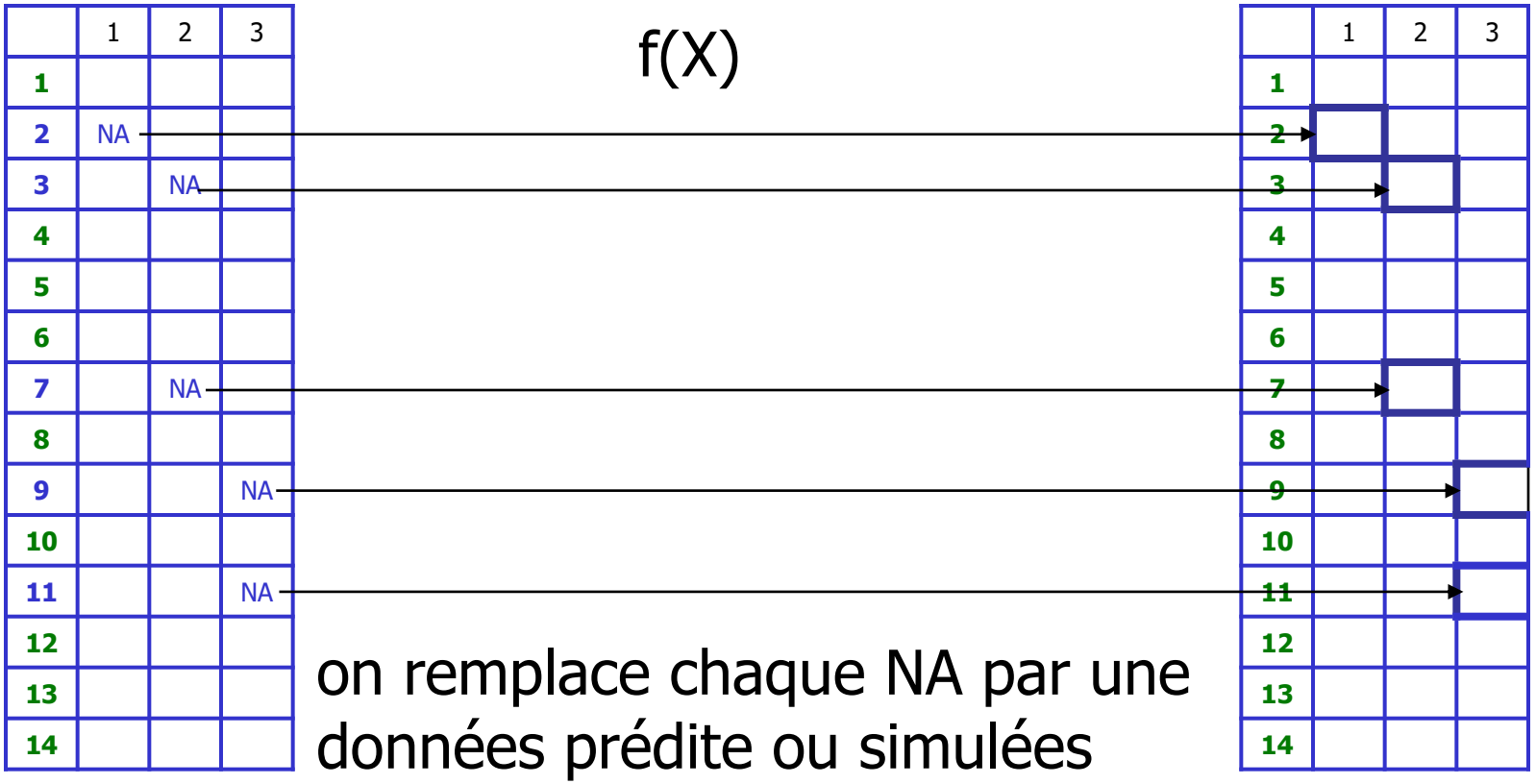
Exemple : Risque prématurité

(Chavance-Manfredi, 2000)

	Antécédent obstétrical pathologique	
	Log(ORa)	ET(Ora)
Données complètes	1,46	0,12
Indicateur de données manquantes	1,36	0,11
Analyse pondérée	1,34	0,12

Imputation Simple (1)

Au lieu de travailler sur …



on remplace chaque NA par une données prédite ou simulées

et l'analyse portera sur tous les enregistrements

Imputation Simple (2)

- 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

IS : Dernière Observation

- 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

	X_Temps0	X_Temps1	X_Temps2	X_Temps3	X_Temps4	X_Temps5
1	18	20	NA	15	21	NA
2	24	NA	NA	NA	15	NA

IS : Hot-Deck et Cold-Deck (1)

- Hot-Deck
  - ✓ La valeur manquante est remplacée par une valeur observée chez un individu ayant les mêmes caractéristiques
- Cold-Deck
  - ✓ La valeur manquante est remplacée par une valeur observée chez un individu ayant les mêmes caractéristiques, mais provenant d'une autre source d'information
- mêmes caractéristiques ⇔ « plus proche voisin »
  - ✓ Fonction de distance basée sur une ou plusieurs variables auxiliaires
  - ✓ Quelle fonction de distance ?

IS : Hot-Deck et Cold-Deck (2)

	X1	X2	X3	X4
1	1	3	1	2
2	1	NA	1	5
3	2	3	NA	2
4	NA	3	NA	5

IS : par la Moyenne

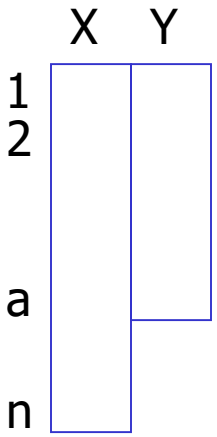
- Remplacement d'une valeur manquante par la moyenne des mesures disponibles
- La même pour toutes les NA d'une même variable
- Estimations non biaisées si les données sont MCAR

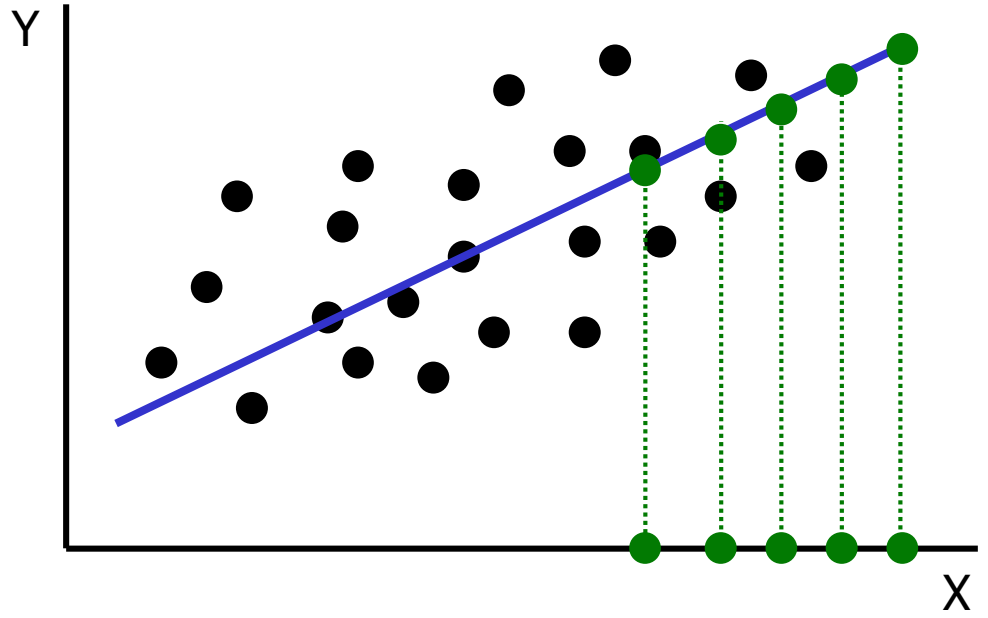
IS : par un modèle de Régression (1)

- 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 (2)

- Exemple : régression linéaire simple





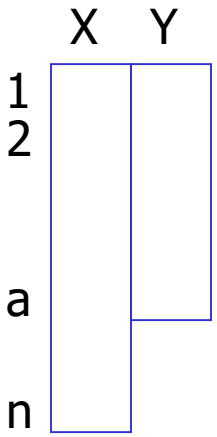
Modélisation sur les cas complets $y_i = \hat{\alpha} + \hat{\beta} x_i \quad i = 1,\ldots,a$

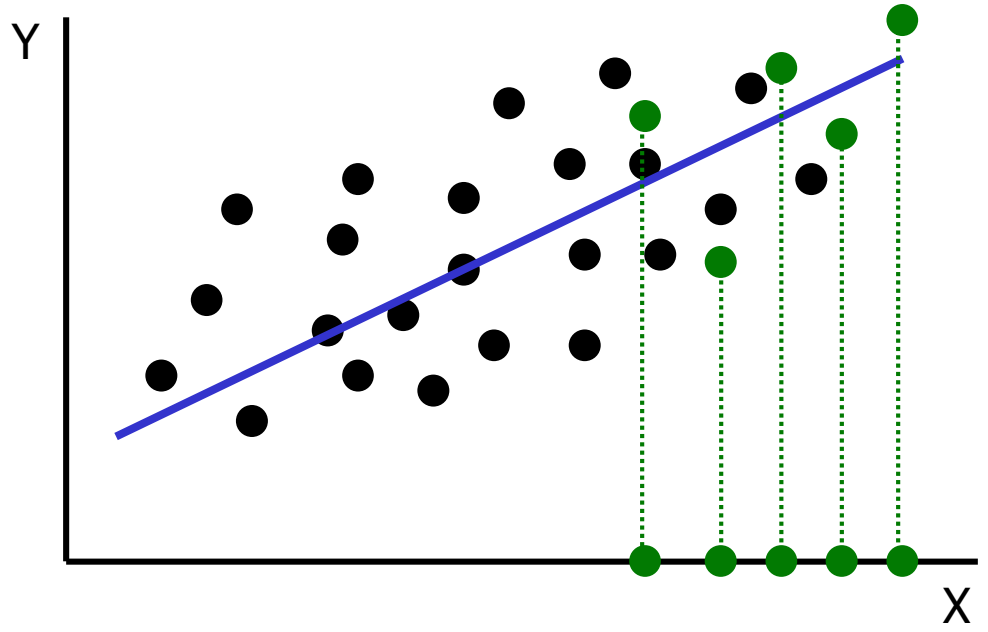
Imputation par la prédiction du modèle de régression

$$y_i^* = \hat{\alpha} + \hat{\beta} x_i \quad i = a+1,\ldots,n$$

IS : par un modèle de Régression (3)

- Exemple : régression linéaire simple avec résidus aléatoires





Modélisation sur les cas complets $y_i = \hat{\alpha} + \hat{\beta} x_i \quad i = 1,\ldots,a \;\; et \;\; \hat{\sigma}^2$

Imputation par la prédiction du modèle de régression

$$y_i^* = \hat{\alpha} + \hat{\beta} x_i + e_i \quad i = a+1,\ldots,n$$

$$\text{avec} \quad e_i \sim N\left(0, \hat{\sigma}^2\right)$$

Exemple : Risque prématurité

(Chavance-Manfredi, 2000)

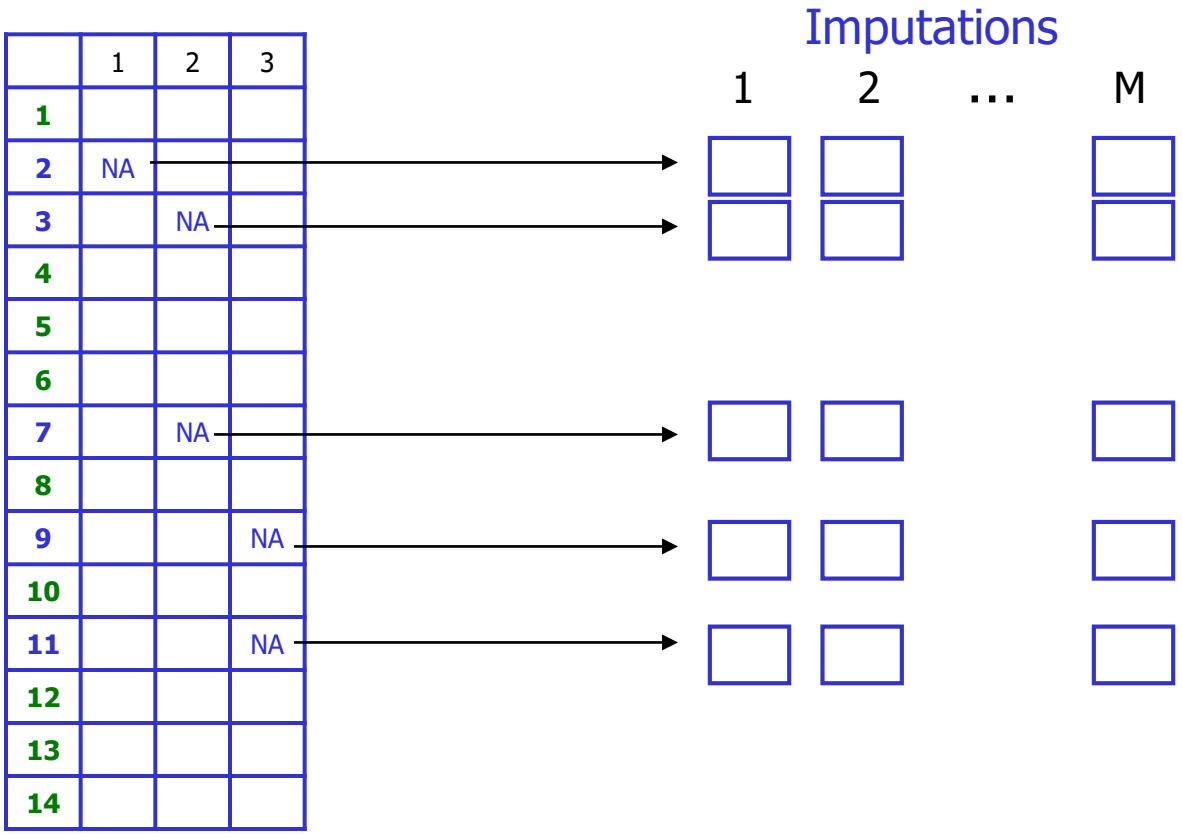
	Antécédent obstétrical pathologique	
	Log(ORa)	ET(Ora)
Données complètes	1,46	0,12
Indicateur de données manquantes	1,36	0,11
Analyse pondérée	1,34	0,12
Imputation simple	1,34	0,11

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

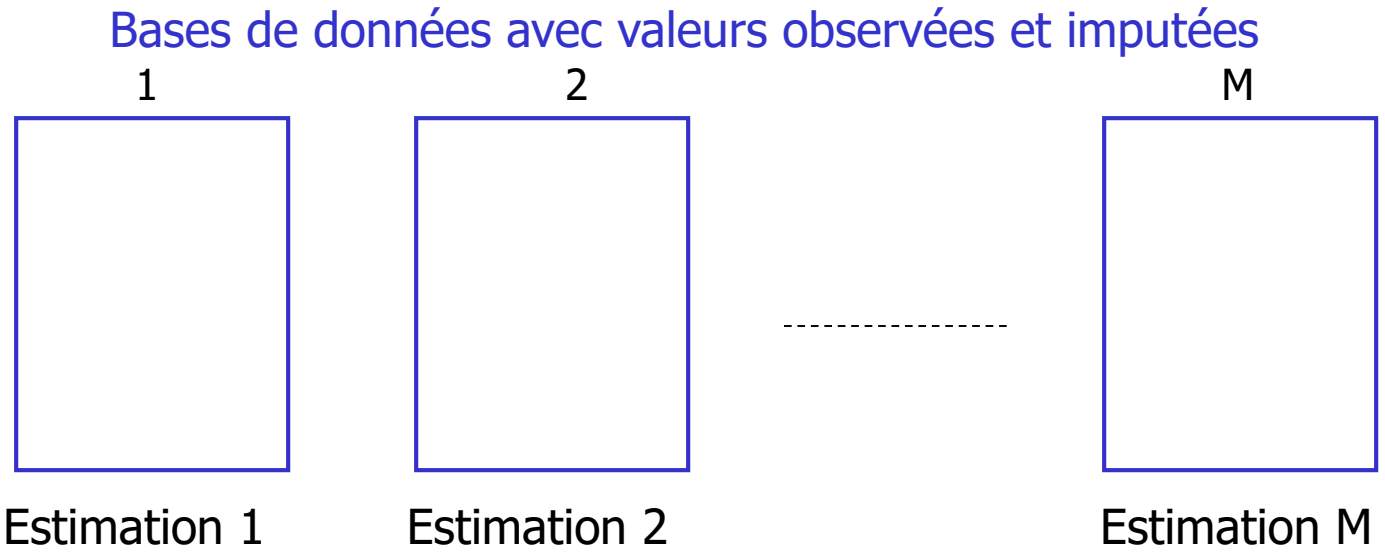
Imputation Multiple : les Étapes (1)

- Remplacer chaque valeur manquante par M > 1 valeurs tirées d'une distribution appropriée



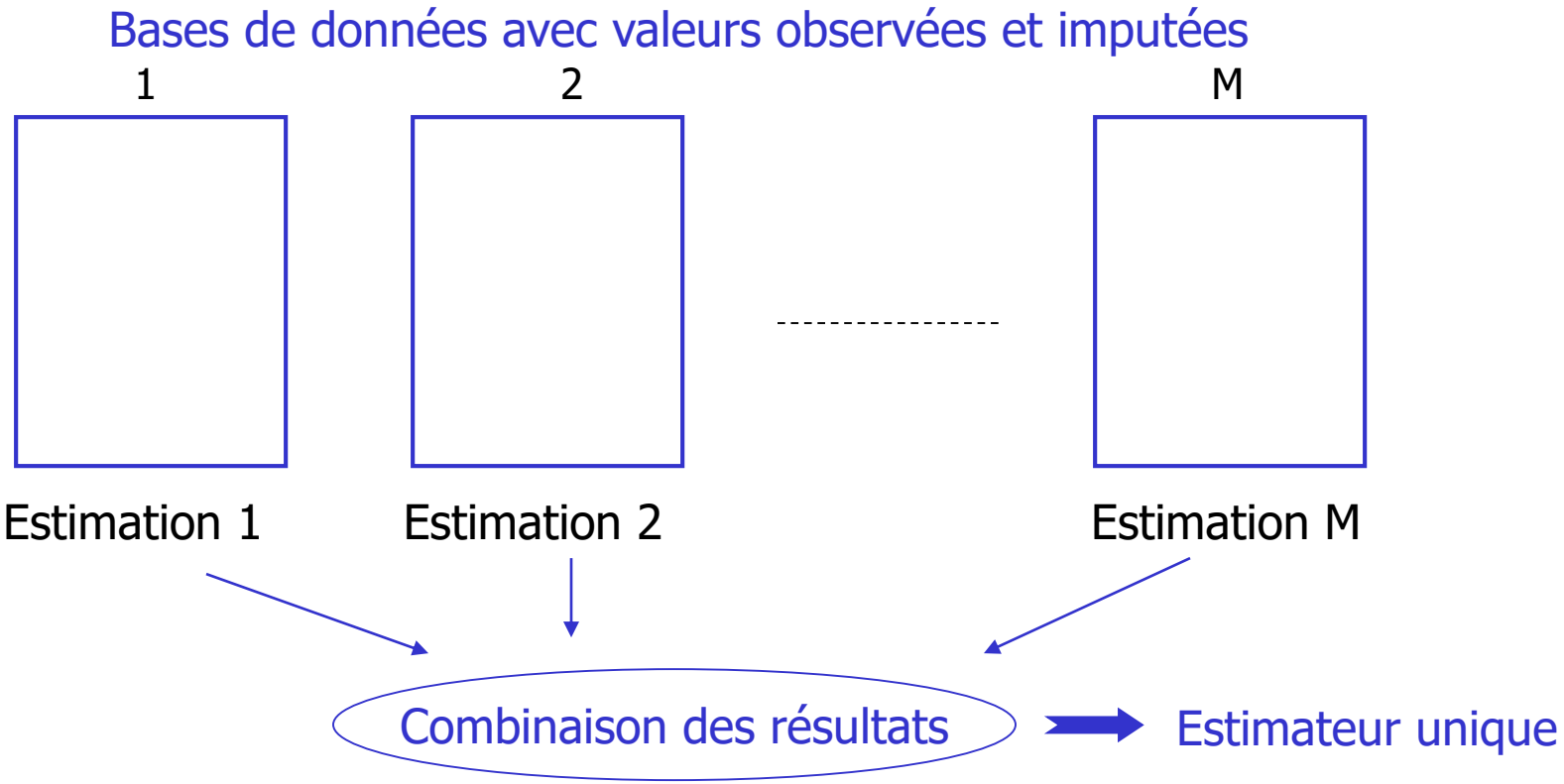
Imputation Multiple : les Étapes (2)

- Analyses indépendantes et avec la même méthode standard des M > 1 bases de données complètes



Imputation Multiple : les Étapes (3)

- Combiner les résultats des analyses afin de refléter la variabilité supplémentaire due aux données manquantes



IM : Commentaires / Remarques (1)

- Approche générale
- Un ensemble d'imputation peut servir pour plusieurs analyses
- Les résultats définitifs incorporent l'incertitude des non réponses
- Très efficace même pour des petites valeurs de M (> 3)
- Hypothèse d'un processus d'observation MAR
- Les distributions prédictives des données manquantes sont très compliquées

IM : Commentaires / Remarques (2)

- IM possibles grâce aux méthodes
  - ✓ Algorithme EM (Expectation – Maximisation)
    - Procédure itérative en 2 étapes
      - Calcul de l'espérance : identification de la distribution des données manquantes en fonction des données observées et des variables explicatives
      - Maximisation : remplace les données manquantes par les valeurs attendues
    - Itération jusqu'à l'obtention d'une stabilisation dans les valeurs attendues
  - ✓ Méthodes de calcul MCMC (Monte Carlo Markov Chain)
    - Création d'une chaîne de Markov qui converge vers la distribution prédictive a posteriori des données manquantes

Exemple : Risque prématurité

(Chavance-Manfredi, 2000)

	Antécédent obstétrical pathologique	
	Log(ORa)	ET(Ora)
Données complètes	1,46	0,12
Indicateur de données manquantes	1,36	0,11
Analyse pondérée	1,34	0,12
Imputation simple	1,34	0,11
Imputation multiple (M =3)	1,34	0,11
Imputation multiple (M =30)	1,34	0,11
Imputation multiple (M =100)	1,34	0,11

Une Synthèse (1)

- Analyses préliminaires des données
  - ✓ Pattern des données manquantes
  - ✓ % de NA pour chaque variables
  - ✓ % de NA par sujets
  - ✓ % de NA total
  - ✓ Corrélation entre les variables
  - ✓ Autres ?

Une Synthèse (2)

- Analyse des données complètes
  - ✓ Méthode la moins satisfaisante en terme de biais et de précision
- Indicateur de données manquantes
  - ✓ Plus précis, permet d'identifier certains problèmes de biais, ne les traite pas de façon pleinement satisfaisante
- Analyse pondérée
  - ✓ Corrige le biais éventuel de données MAR, mais augmente l'imprécision

Une Synthèse (3)

- Imputation simple
  - ✓ Exige souvent un processus d'observation MCAR
- Imputation multiple
  - ✓ Prend en compte simultanément les problèmes de biais et de précision
  - ✓ Flexible
  - ✓ Adaptée pour des données qualitatives et quantitatives
  - ✓ Utilisable pour différents type d'analyse (régression logistique, …)

Imputation Multiple : Logiciels

- SOLAS
  - ✓ http://www.statsol.ie/html/solas/solas_home.html
- SAS PROC MI et PROC MIANALYZE
  - ✓ http://support.sas.com/rnd/app/da/new/dami.html
- S-Plus
  - ✓ http://www.insightful.com/
- MICE* pour R et S-Plus
  - ✓ http://www.multiple-imputation.com
- NORM*
  - ✓ http://www.stat.psu.edu/~jls/misoftwa.html

* Gratuit

Imputation Multiple : MICE

- Package MICE à télécharger et à installer dans R
  - ✓ http://cran.r-project.org/ dans Software, Packages
- Avant son utilisation dans R
  - ✓ Il faut le charger dans R
  - ✓ Menu « Packages », « Charger le package… », sélectionner « `mice` »
- Contient le fichier `nhanes`
  - ✓ 25 sujets
  - ✓ 4 variables
    - `age` : groupes d'âge : 1=20-39, 2=40-59, 3=60+
    - `bmi` : index de masse corporelle (kg/m^2)
    - `hyp` : hypertension : 1=non, 2=oui
    - `chl` : cholestérol total (mg/dL)

Exemple `nhanes` (1)

- Importation des données
- Visualisation des données pour les 5 premières lignes

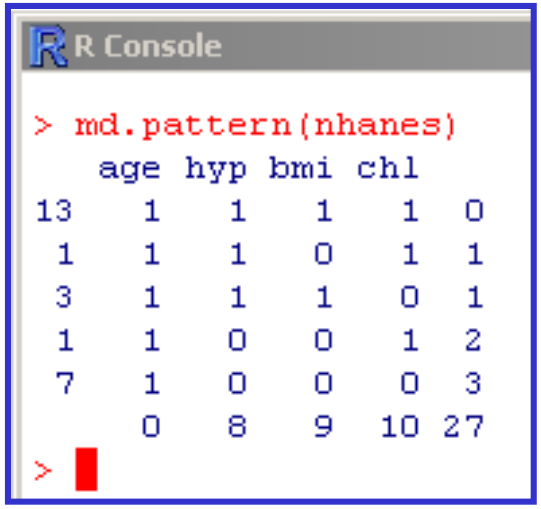
- Proportion de données NA par variables
- Proportion de données NA par individus

```
R Console
> data(nhanes)
> nhanes[1:5 , ]
  age  bmi hyp chl
1   1   NA  NA  NA
2   2 22.7   1 187
3   1   NA   1 187
4   3   NA  NA  NA
5   1 20.4   1 113
> colSums(is.na(nhanes))/nrow(nhanes)
 age  bmi  hyp  chl
0.00 0.36 0.32 0.40
> rowSums(is.na(nhanes))/ncol(nhanes)
   1    2    3    4    5    6    7    8
0.75 0.00 0.25 0.75 0.00 0.50 0.00 0.00
   9   10   11   12   13   14   15   16
0.00 0.75 0.75 0.75 0.00 0.00 0.25 0.75
  17   18   19   20   21   22   23   24
0.00 0.00 0.00 0.25 0.75 0.00 0.00 0.25
  25
0.00
>
```

Exemple `nhanes` (2)

- Pattern des données manquantes

13 lignes sont complètes →
3 lignes où seule `chl` est NA →
Nombre total de NA = 27 →



Exemple `nhanes` (3)

- La fonction `mice` (multivariate imputation by chained equations)
- Arguments
  - `data` : data frame, matrice des données (NA)
  - `m` : nombre d'imputations multiples (m=5 *)
  - `imputationMethod` : méthode utilisée pour les imputations
    - `norm` : régression linéaire bayésienne (numeric)
    - `pmm` *: moyenne prédite par « appariement » (numeric)
    - `mean` : moyenne marginale (numeric)
    - `logreg` *: régression logistique (2 catégories)
    - `logreg2` : régression logistique (minimisation directe) (≥ 2 catégories)
    - `polyreg` *: régression logistique multinomiale (≥ 2 catégories)
    - `lda` : analyse discriminante linéaire (≥ 2 catégories)
    - `sample` : échantillonnage aléatoire à partir des données observées

* Valeurs par défaut

Exemple `nhanes` (4)

> Algorithme d'imputation

Valeur par défaut →

Récapitulatif des NA →

Méthodes par défaut →

Ordre dans lequel les variables sont utilisées pour l'imputation →

Indicatrice des variables utilisées pour chaque variable avec NA →

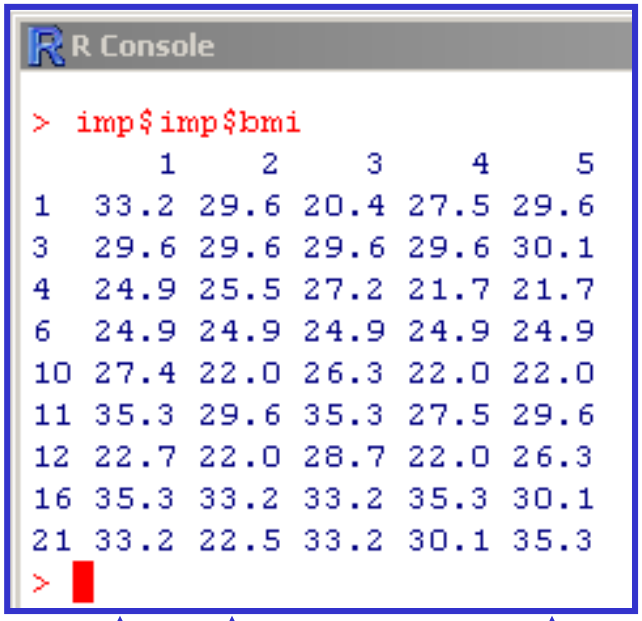
R Console

```
> imp <- mice(nhanes)
> imp
Multiply imputed data set
Call:
mice(data = nhanes)
Number of multiple imputations:  5
Missing cells per column:
age bmi hyp chl
  0   9   8  10
Imputation methods:
  age   bmi   hyp   chl
   "" "pmm" "pmm" "pmm"
VisitSequence:
bmi hyp chl
  2   3   4
PredictorMatrix:
    age bmi hyp chl
age   0   0   0   0
bmi   1   0   1   1
hyp   1   1   0   1
chl   1   1   1   0
Random generator seed value:  NA
>
```

Exemple `nhanes` (5)

> Données imputées pour `bmi`

Lignes où `bmi` est NA



Valeurs 1ère imputation

2ème imputation

...

5ème imputation

Exemple `nhanes` (6)

- Visualisation des 5 premières lignes du premier jeux de données complètes

```
R Console

> complete(imp, 1)[1:5 , ]
  age  bmi hyp chl
1   1 33.2   1 229
2   2 22.7   1 187
3   1 29.6   1 187
4   3 24.9   1 184
5   1 20.4   1 113
>
```

⋮

- Visualisation des 5 premières lignes du cinquième jeux de données complètes

```
R Console

> complete(imp, 5)[1:5 , ]
  age  bmi hyp chl
1   1 29.6   1 187
2   2 22.7   1 187
3   1 30.1   1 187
4   3 21.7   1 206
5   1 20.4   1 113
>
```

Exemple `nhanes` (7)

- Analyse complète
  - ✓ Régression linéaire de `chl` sur `age` et `hyp`

```
R Console
① > fit <- lm.mids(chl ~ age + hyp, imp)
② > pool(fit)
Call: pool(object = fit)

③ Pooled coefficients:
(Intercept)          age          hyp
  128.11019    16.88824    33.15435

④ Fraction of information about the coefficients missing due to nonresponse:
(Intercept)          age          hyp
  0.1638523    0.2561456    0.2518272
>
```

① Régression linéaire sur des données imputées multiples
② Pour obtenir les résultats uniques des m analyses complètes
③ Résultat final de la régression linéaire
④ Fraction de l'information due à la non-réponse

Exemple `nhanes` (8)

- Analyse complète
  - ✓ Régression linéaire de `chl` sur `age` et `hyp`

```
R Console

> summary(pool(fit))
                   est        se         t        df     Pr(>|t|)
(Intercept) 128.11019 27.69687 4.625439 16.78156 0.0002494190
age          16.88824 12.19378 1.384988 13.35101 0.1887631610
hyp          33.15435 24.99010 1.326700 13.50286 0.2066086715
                lo 95     hi 95 missing       fmi
(Intercept)  69.616901 186.60348      NA 0.1638523
age          -9.384586  43.16107       0 0.2561456
hyp         -20.629804  86.93851       8 0.2518272
>
```

Résultats plus généraux de l'analyse de régression linéaire sur des données imputées multiples

Références & Liens Utiles

- Chavance M, Manfredi R. Modélisation d'observations incomplètes. *Rev Epidém et Santé Publ* 2000; 48:389-400.
- Schafer JL. *Analysis of incomplete multivariate data*. Chapman & Hall, 2000.
- Horton NJ, Stuart RL. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician* 2001;55:244-254. (http://www.biostat.harvard.edu/~horton/tasimpute.pdf)
- van Buuren S, Oudshoorn CGM. *Multivariate imputation by chained equations*. MICE V1.0 User's manual.

  (http://web.inter.nl.net/users/S.van.Buuren/mi/docs/Manual.pdf)
- Multiple imputation online : http://www.multiple-imputation.com
- The multiple imputation FAQ page : http://www.stat.psu.edu/~jls/mifaq.html