Données manquantes

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Données manquantes

Plusieurs champs d'une base de donnée peuvent être manquants

- non-réponse
- valeurs erronées (erreur d'encodage)
- perte de suivi et censure
- plusieurs versions de formulaires (question optionnelles)

Pourquoi s'en préoccuper?

La plupart des procédures ne gèrent que les cas complets (toute observation avec des valeurs manquantes est éliminée).

Les données manquantes réduisent l'information disponible.

Sans traitement adéquat, les estimations seront **biaisées**.

Références

- van Buuren, S. (2018). *Flexible imputation of missing data*, CRC Press, 2e édition.
- Little, R. et D. Rubin (2019). *Statistical Analysis with Missing Data*, Wiley, 3e édition
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall / CRC.

Les valeurs manquantes dans un contexte de prédictions sont couvertes dans le cours MATH 60600.

Classification des données manquantes (MCAR)

Cas 1: Données manquantes de façon complètement aléatoire (*missing completely at random*)

La probabilité que la valeur soit manquante ne dépend ni de la valeur, ni de celles des autres variables.

Exemple: questionnaire trop long, la personne ne répond pas à tout (sans lien avec les questions posées).

Hypothèse souvent irréaliste en pratique.

Classification des données manquantes (MAR)

Cas 2: données manquantes de façon aléatoire (*missing at random*): la probabilité que la valeur soit manquante ne dépend pas de la valeur *une fois qu'on a contrôlé pour les autres variables*.

Exemple: les hommes sont plus susceptibles dans l'ensemble de divulguer leur âge que les femmes.

Classification des données manquantes (MNAR)

Cas 3: données manquantes de façon non-aléatoire (*missing not at random*): la probabilité que la mesure soit manquante dépend de la valeur elle-même, pas déterminable avec d'autres variables

Exemple: une personne transgenre ne répond pas à la question genre (si seulement deux choix, homme/femme) et aucune autre question ne se rattache au genre ou à l'identité sexuelle.

Comment déterminer le type de données manquantes

Une personne ne divulgue pas son salaire? Données manquante de manière aléatoire ou non aléatoire?

Hypothèse pas testable, dépend du contexte et des variables auxiliaires disponibles.

Valeur logique

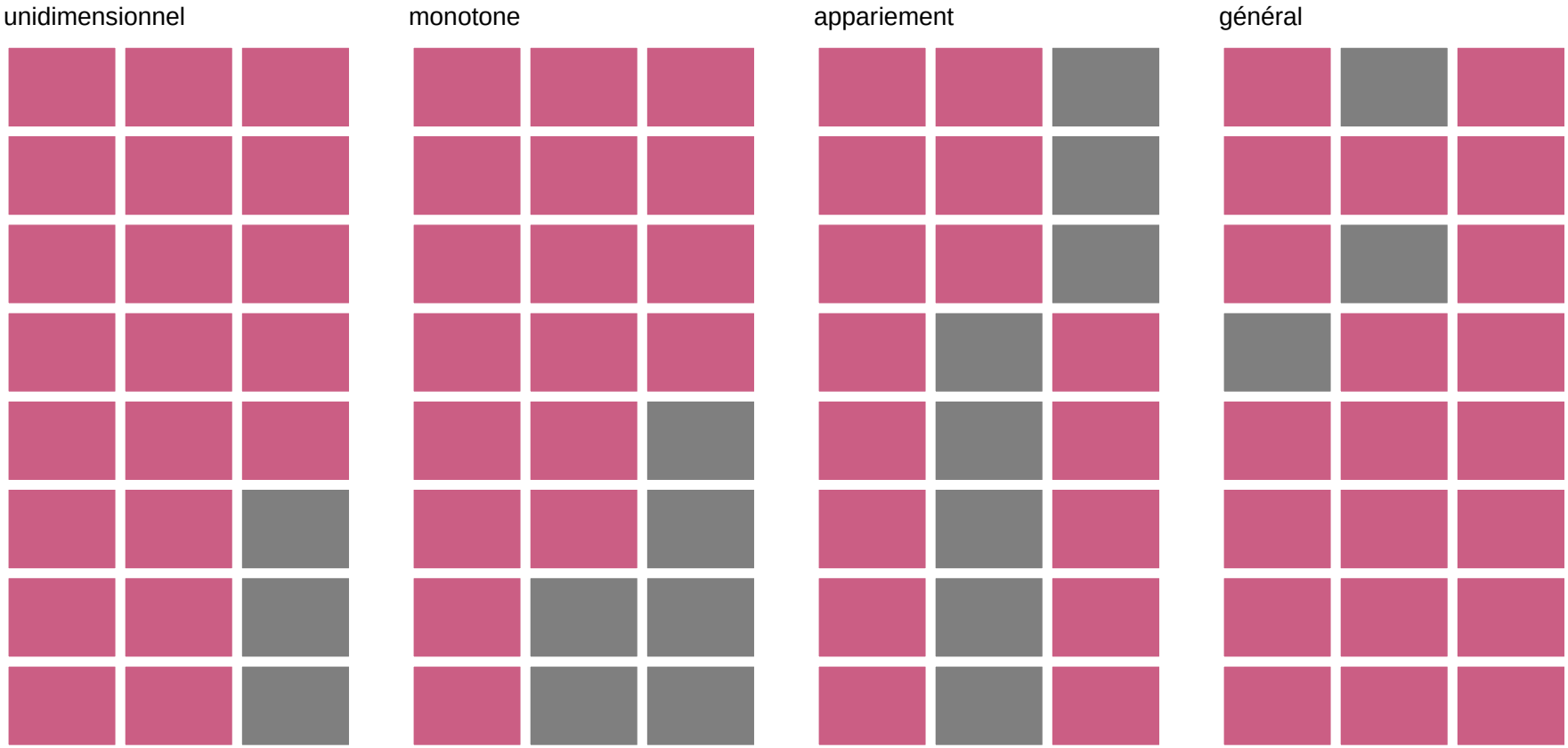
Les données manquantes ont souvent une valeur logique:

- un client qui n'a pas de carte de crédit a un solde de 0!

D'où l'importance des validations d'usage et du nettoyage préliminaire de la base de données.

Types de schémas de données manquantes

Matrice $n \times p$ (observations en lignes, variables en colonnes).



Les cases grises représentent des valeurs manquantes. Illustration adapté de la Figure 4.1 de van Buuren (2022)

Analyse des cas complet

Retirer les observations avec données manquantes pour conserver les cas complets.

- Valide uniquement pour complètement aléatoire.
- On perd de la précision en utilisant moins d'observations.

Méthode par défaut dans les logiciels.

Méthodes d'imputation: imputation simple

Imputation: emplacer les valeurs manquantes par une valeur judicieuse pour *combler les trous*.

> *Le concept d'imputation est à la fois séduisant et dangereux (Dempster et Rubin, 1983)*

Par exemple, on remplacer la valeur manquante par la moyenne (variables continues) ou le mode (variables catégorielles).

Approche pas recommandée (pourquoi?)

Faut-il toujours imputer?

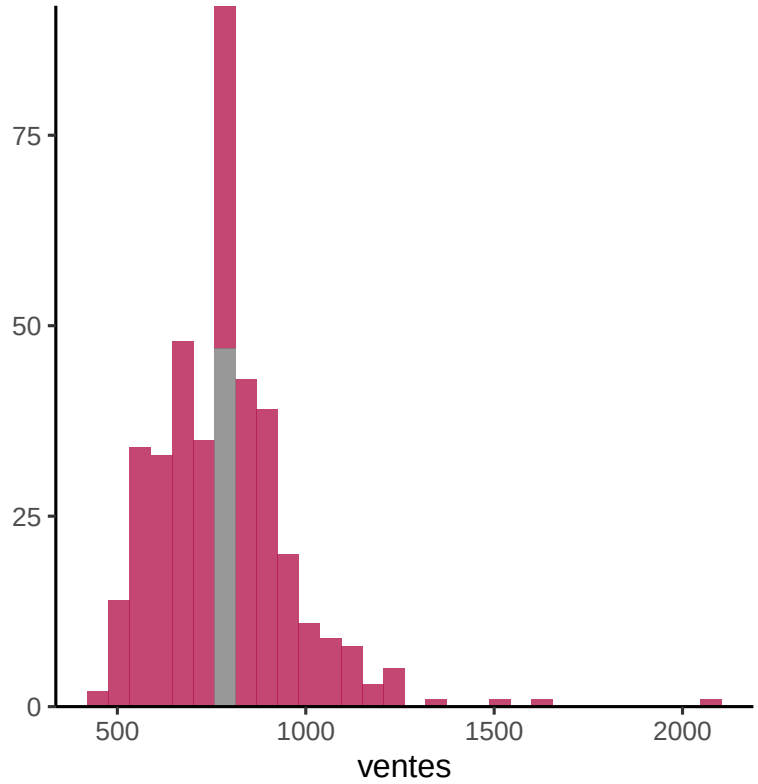
Il faut utiliser son jugement.

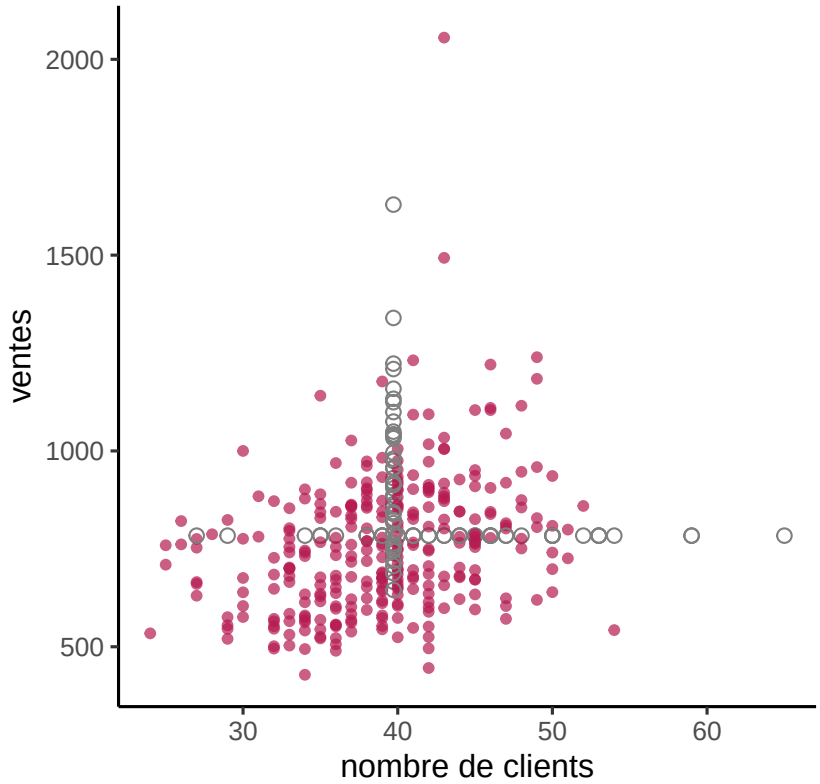
Une observation imputée ne remplacera jamais une vraie observation.

- Si la proportion d'observations manquantes est petite (moins de 5%), on pourrait faire une analyse avec les cas complets (et valider au besoin en utilisant l'imputation multiple).
- Si la proportion de valeurs manquantes est 30% et que cette proportion baisse à 3% lorsque vous éliminez quelques variables peu importantes pour votre étude, alors procédez à leur élimination.

Imputation par la moyenne

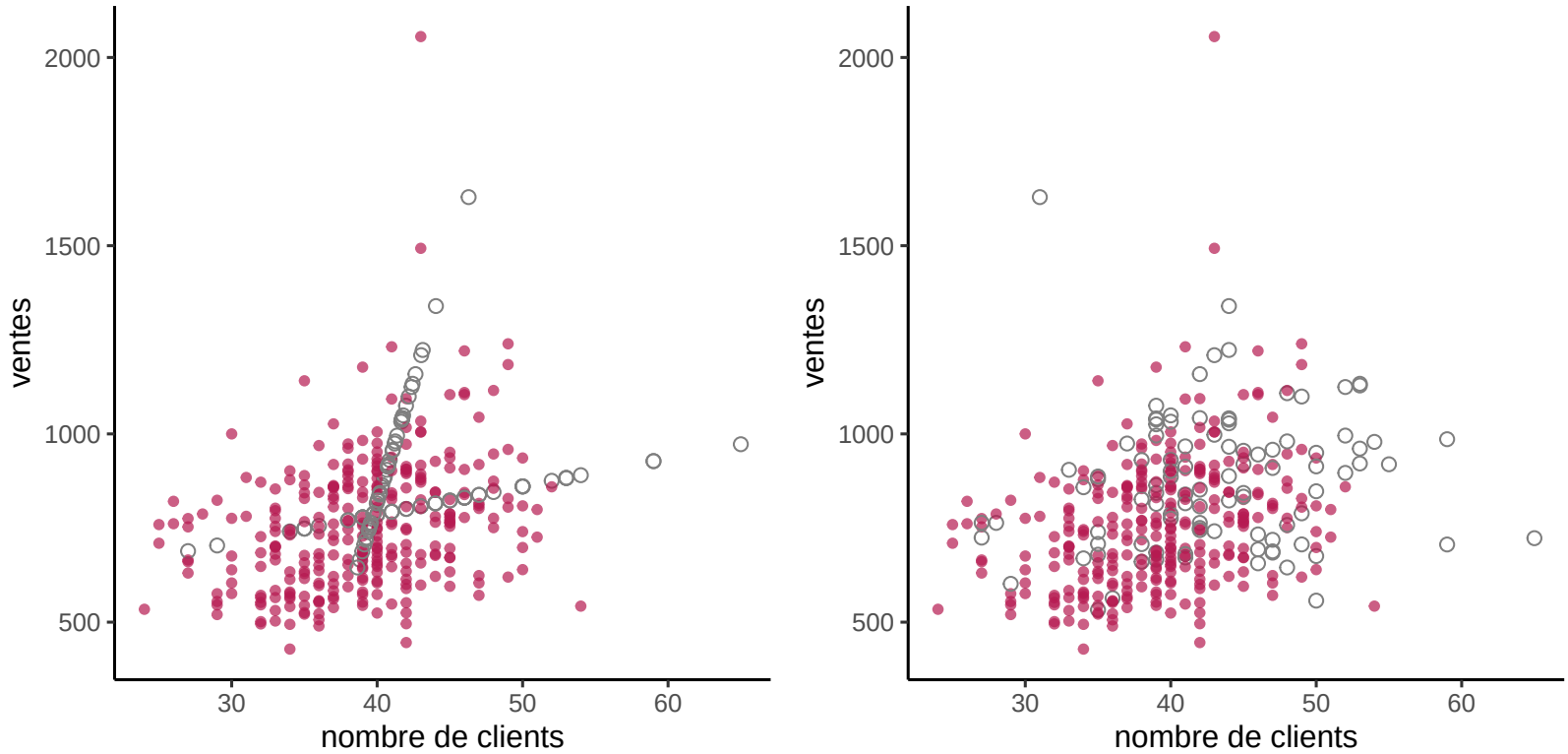
Dilution de la relation (corrélation) entre variables explicatives.
Réduction de la variabilité.





Modèles prédictifs

L'imputation par régression (gauche) mène à une sous-estimation de l'incertitude en raison de l'augmentation de la corrélation, contrairement à l'imputation aléatoire (droite).



Imputation aléatoire

Considérons le cas d'une régression logistique pour une variable explicative binaire.

Plutôt que d'assigner à la classe la plus probable, une prédiction aléatoire simule une variable 0/1 avec probabilité $(1 - \hat{p}_i, \hat{p}_i)$.

```
pred <- 0.3 #probabilité de succès
rbinom(n = 15, size = 1, prob = pred)
```

```
 [1] 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0
```

Problèmes de l'imputation simple

On ne tient pas compte du fait que des valeurs ont été remplacées (on fait comme si c'était de vraies observations).

On sous-évalue encore une fois la **variabilité** des données

- les écarts-type des estimations sont trop petits.

Inspection des valeurs manquantes

Il est donc nécessaire d'examiner la configuration des valeurs manquantes avant de faire quoi que ce soit.

```
data(manquantes, package = 'hecmulti')
summary(manquantes)
# Pourcentage de valeurs manquantes
apply(manquantes, 2, function(x){mean(is.na(x))})
# Voir les configurations de valeurs manquantes
md.pattern(manquantes) # graphique diapo suivante
```

Table 1: Nombre et pourcentage de valeurs manquantes par variable.

	x1	x2	x3	x4	x5	x6	y
nombre	192	49	0	184	0	0	0
pourcentage	38.4	9.8	0	36.8	0	0	0

Configuration des valeurs manquantes

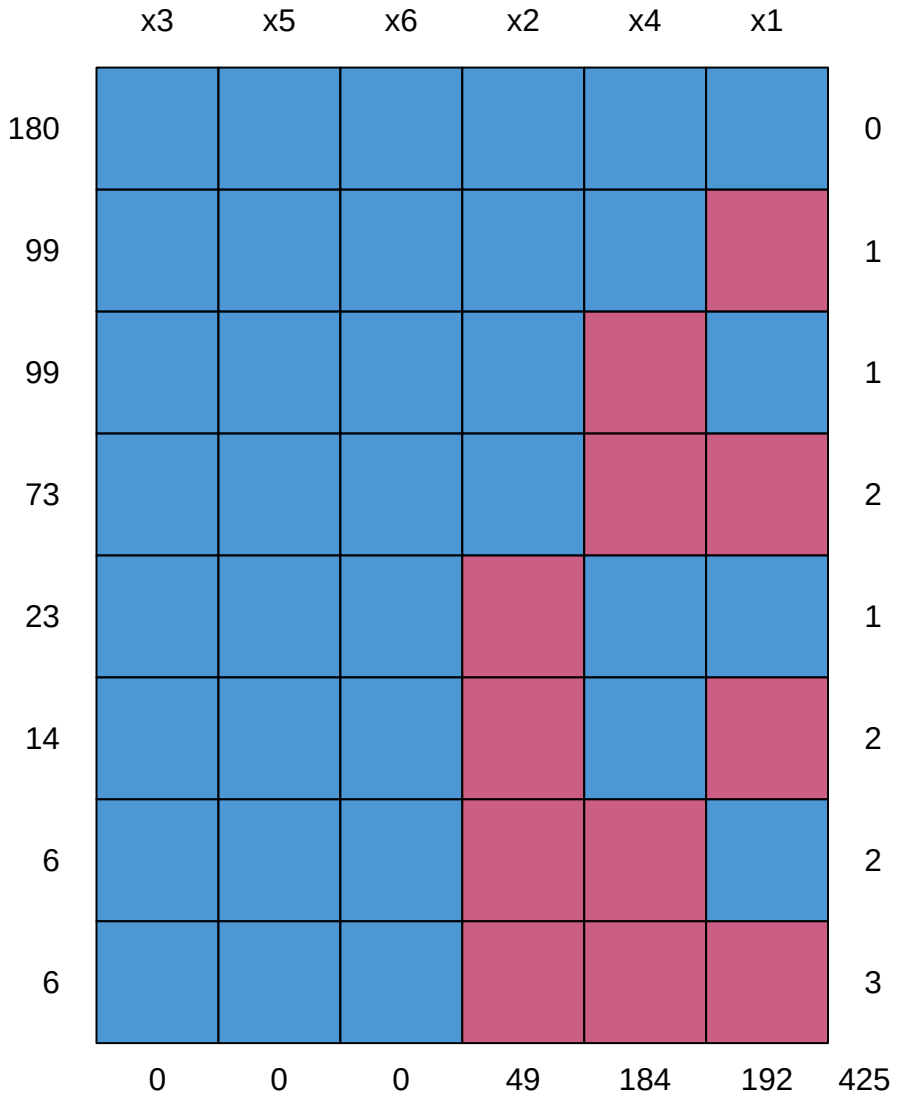
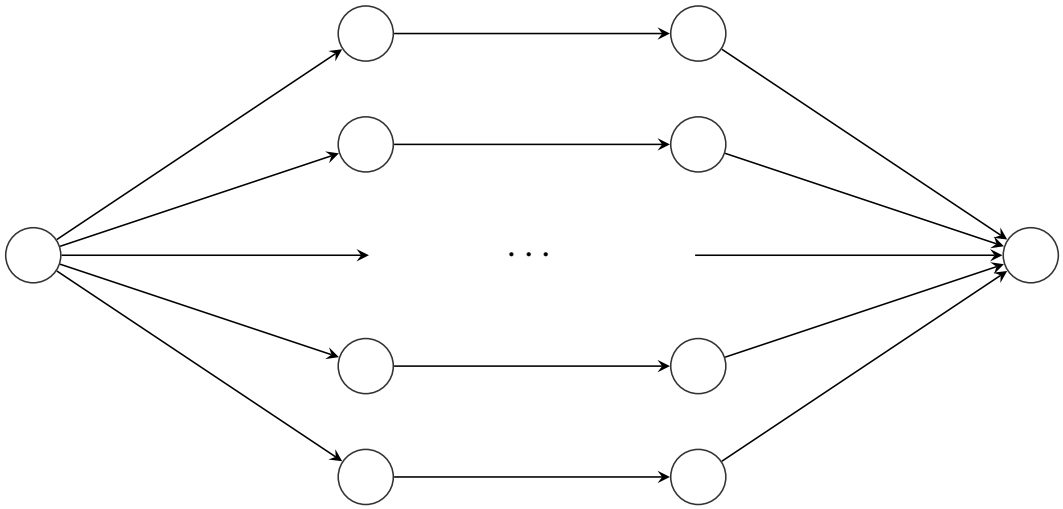


Figure 1: Configurations des valeurs manquantes pour `manquantes`.

Imputation multiple

Valides pour les données manquantes de manière (complètement) aléatoires.

1. Procéder à plusieurs imputations **aléatoires** pour obtenir un échantillon complet (`mice`)
2. Ajuster le modèle d'intérêt avec chaque échantillon (`with`). 3. Combiner les résultats obtenus (`pool` et `summary`)



Combinaison des résultats

Considérons un seul paramètre θ (ex: coefficient d'une régression) et supposons qu'on procède à K imputations.

On estime les paramètres du modèle séparément pour chacun des K ensembles de données imputés, disons

- $\widehat{\theta}_k$ pour l'estimation du paramètre θ dans l'échantillon k et
- $\widehat{\sigma}_k^2 = \mathsf{Va}(\widehat{\theta}_k)$ pour l'estimation de la variance de $\widehat{\theta}_k$.

Estimation du paramètre moyen

L'estimation finale de θ, dénotée $\hat{\theta}$, est obtenue tout simplement en faisant la moyenne des estimations de tous les modèles, c'est-à-dire,

$$\hat{\theta} = \frac{\hat{\theta}_1 + \cdots + \hat{\theta}_K}{K}.$$

Estimation des erreurs-types

Une estimation ajustée de la variance de $\hat{\theta}$ est

$$\mathsf{Va}(\hat{\theta}) = W + \frac{K+1}{K} B,$$

où

$$W = \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2 = \frac{\hat{\sigma}_1^2 + \cdots + \hat{\sigma}_K^2}{K},$$

$$B = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2.$$

- W est la moyenne des variances (variance intra-groupe) et
- B la variance des moyennes (variance inter-groupe).

Imputation multiple

Des formules analogues existent pour les degrés de liberté, les valeurs-p, etc. ainsi que pour la cas multidimensionnel (plusieurs paramètres).

Si on procédait à une seule imputation (même en ajoutant une part d'aléatoire pour essayer de reproduire la variabilité des données), on ne serait pas en mesure d'estimer la variance inter-groupe de l'estimateur.

On peut estimer la fraction de l'information manquante sur θ avec $(1 + 1/K)B/\mathsf{Va}(\hat{\theta})$.

Imputation multiple par équations chaînées (MICE)

Avec p variables $X_1, \dots, X_p$, spécifier un ensemble de modèles **conditionnels** pour chaque variable X_j en fonction de

- toutes les autres variables, X_{-j}
- les valeurs observées pour cette variable, $X_{j,\text{obs}}$

1. Initialisation: remplir les trous avec des données au hasard parmi $X_{j,\text{obs}}$ pour $X_{j,\text{man}}$
2. À l'itération t, pour chaque variable $j = 1, \dots, p$, à tour de rôle:
   a) tirage aléatoire des paramètres $\phi_j^{(t)}$ du modèle pour $X_{j,\text{man}}$ conditionnel à $X_{-j}^{(t-1)}$ et $X_{j,\text{obs}}$
   b) échantillonnage de nouvelles observations $X_{j,\text{man}}^{(t)}$ du modèle avec paramètres $\phi_j^{(t)}$ conditionnel à $X_{-j}^{(t-1)}$ et $X_{j,\text{obs}}$
3. Répéter le cycle

Imputation multiple avec `mice`

```r
library(mice)
# Intensif en calcul, réduire "m" si nécessaire
impdata <- mice(
    data = manquantes,
    # argument "method" pour le modèle
    # dépend du type des variables, par ex.
    # régression logistique pour données binaires
    m = 50, # nombre d'imputations
    seed = 60602, # germe aléatoire
    printFlag = FALSE)
# Extraite une copie (m=1,..., 50) imputée
complete(data = impdata,
         action = 1) #no de la copie
```

Estimation et combinaison avec `mice`

```r
# ajuste le modèle avec les données imputées
adj_im <- with(
  data = impdata,
  expr = glm(y ~ x1 + x2 + x3 + x4 + x5 + x6,
             family = binomial))
# combinaison des résultats
fit <- pool(adj_im)
summary(fit)
```

Résultats

terme	estimation	erreur-type	stat	ddl	valeur-p
(Intercept)	-2.86	1.08	-2.65	327.28	0.009
x12	-0.90	0.58	-1.55	172.63	0.124
x13	-0.53	0.59	-0.90	164.98	0.372
x14	-0.76	0.55	-1.39	241.12	0.166
x15	-0.58	0.64	-0.91	176.76	0.365
x22	-0.03	0.44	-0.07	403.25	0.944
x23	-0.65	0.49	-1.34	353.29	0.181
x24	-2.60	0.65	-4.00	271.66	<1e-04
x25	-1.32	0.72	-1.84	283.49	0.066
x3	1.61	0.31	5.19	300.31	<1e-04
x4	2.56	0.42	6.14	185.05	<1e-04
x5	0.11	0.02	5.27	436.59	<1e-04
x62	-1.41	0.37	-3.83	408.17	1e-04
x63	-2.58	0.41	-6.24	399.50	<1e-04

Récapitulatif

- Les données manquantes réduisent la quantité d'information disponible et augmentent l'incertitude.
- On ne peut **pas** les ignorer (étude des cas complets) sans biaiser les interprétations et réduire la quantité d'information disponible.
- Pour bien capturer l'**incertitude** et ne pas modifier les relations entre variables, il faut utiliser une méthode **aléatoire**.
- Avec l'algorithme MICE, on utilise un modèle conditionnel pour chaque variable à tour de rôle

Récapitulatif

L'imputation multiple est préférée à l'imputation simple car elle permet d'estimer l'incertitude sous-jacente en raison des données manquantes.

- On procède à l'imputation plusieurs fois (avec un modèle conditionel, prédictions différentes chaque fois)
- on crée plusieurs copies
- ajuste le modèle sur chacune et
- combine les résultats

Traitement spécial pour erreurs-type, degrés de liberté, valeurs-p et intervalles de confiance.