Données Manquantes

M2 Stats de la SD, 2024-2025

Hadrien Lorenzo
Aix Marseille Université

Au menu

Introduction

Définition et analyse cas complet

Définition

On dit qu'une donnée est manquante lorsque l'observation d'une variable pour une observation n'est pas accessible

On peut choisir de retirer l'observation associée aux NA mais peu indiqué si :

- petit échantillon (n faible),
- grande quantité de NA.

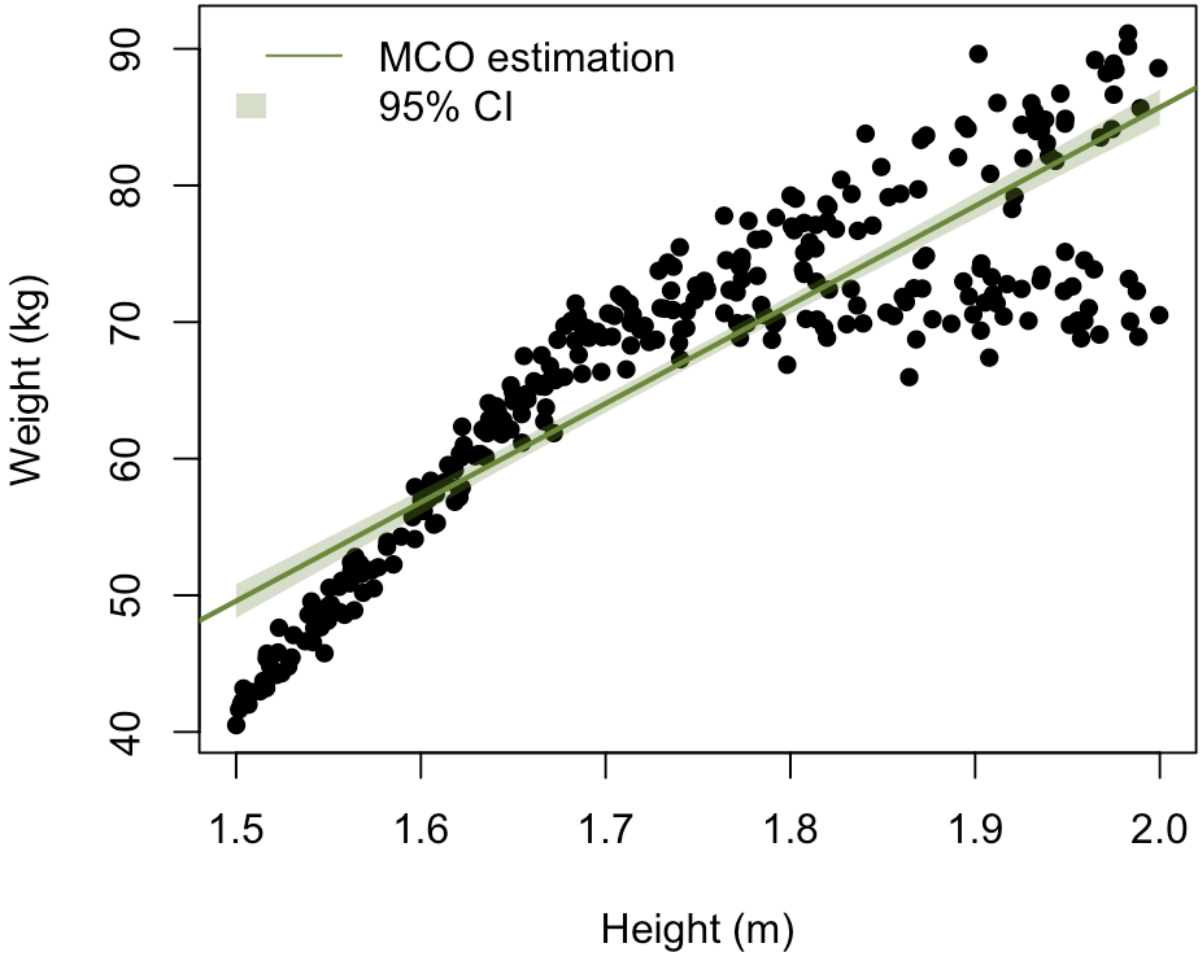
On peut aussi retirer les variables associées aux NA si :

- la structure des données le permet (p important)
- **et** peu de variables sont atteintes par des NA

Le jeu de données résultant est appelé **cas complet**, en pratique il n'est jamais conseillé d'utiliser cette approche… Pourquoi ?

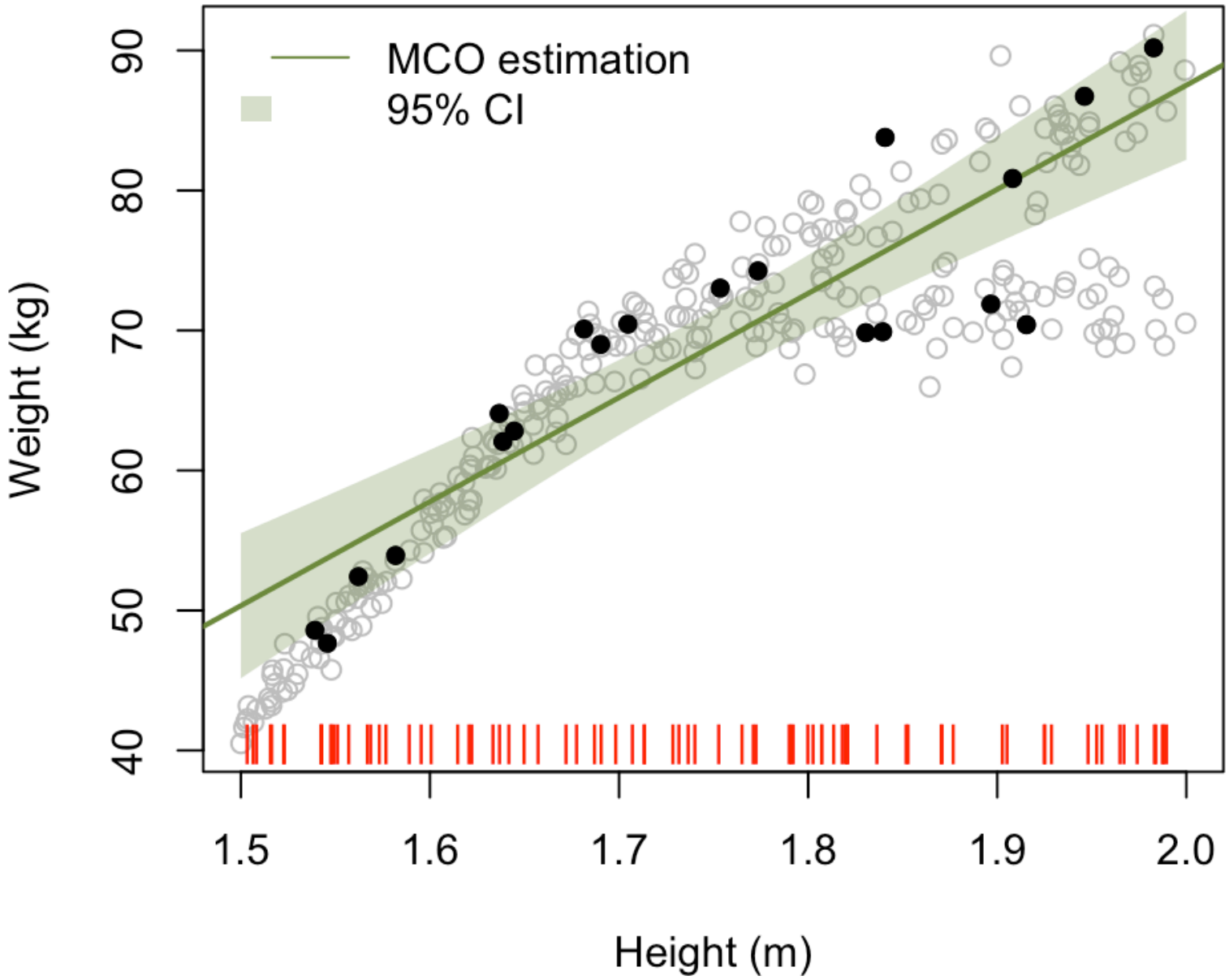
Exemple

Soit le jeu de données : $n = 300$ personnes à qui on a demandé de renseigner leur taille et leur poids. Si tout va bien :

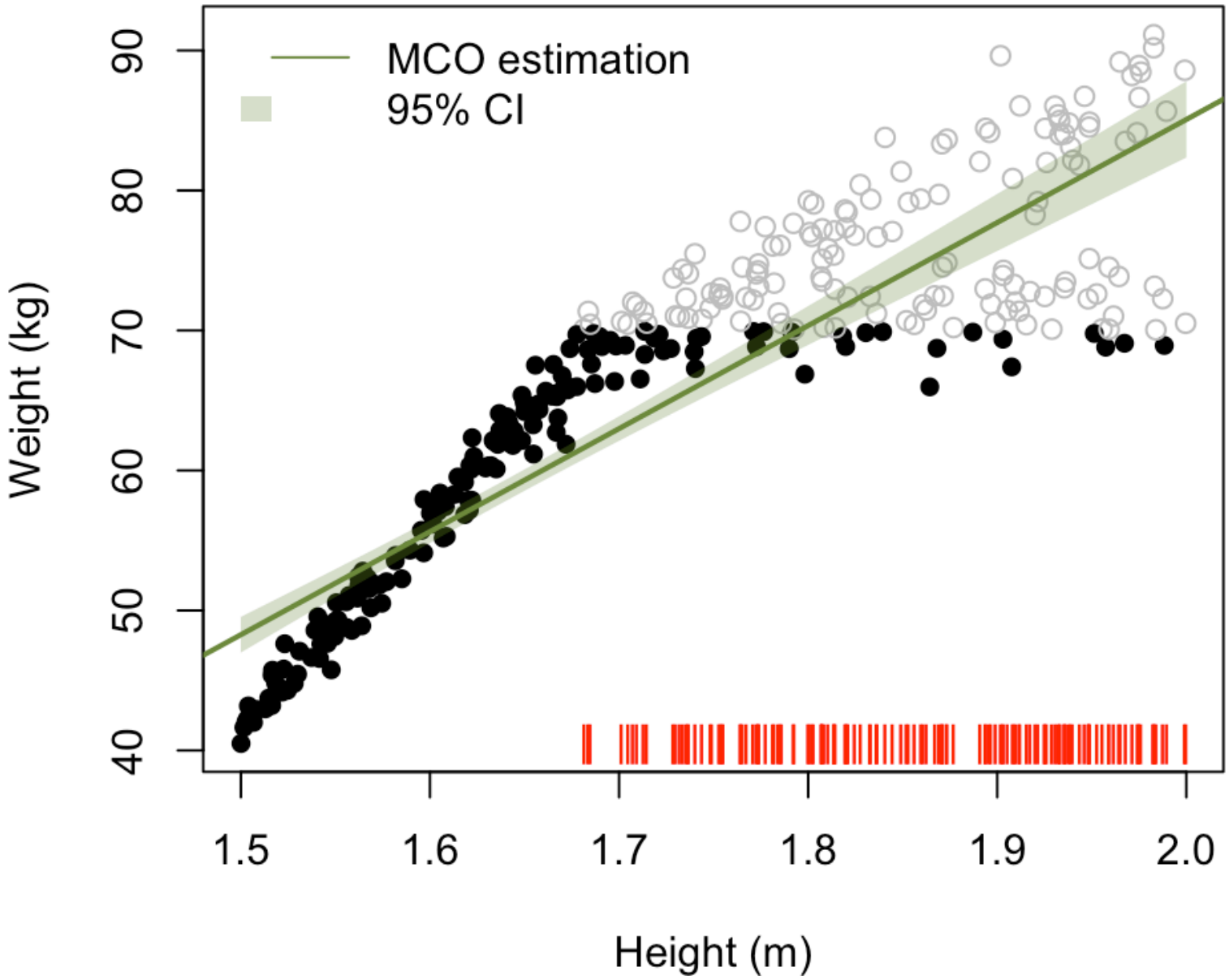


… mais plusieurs problèmes peuvent se produire

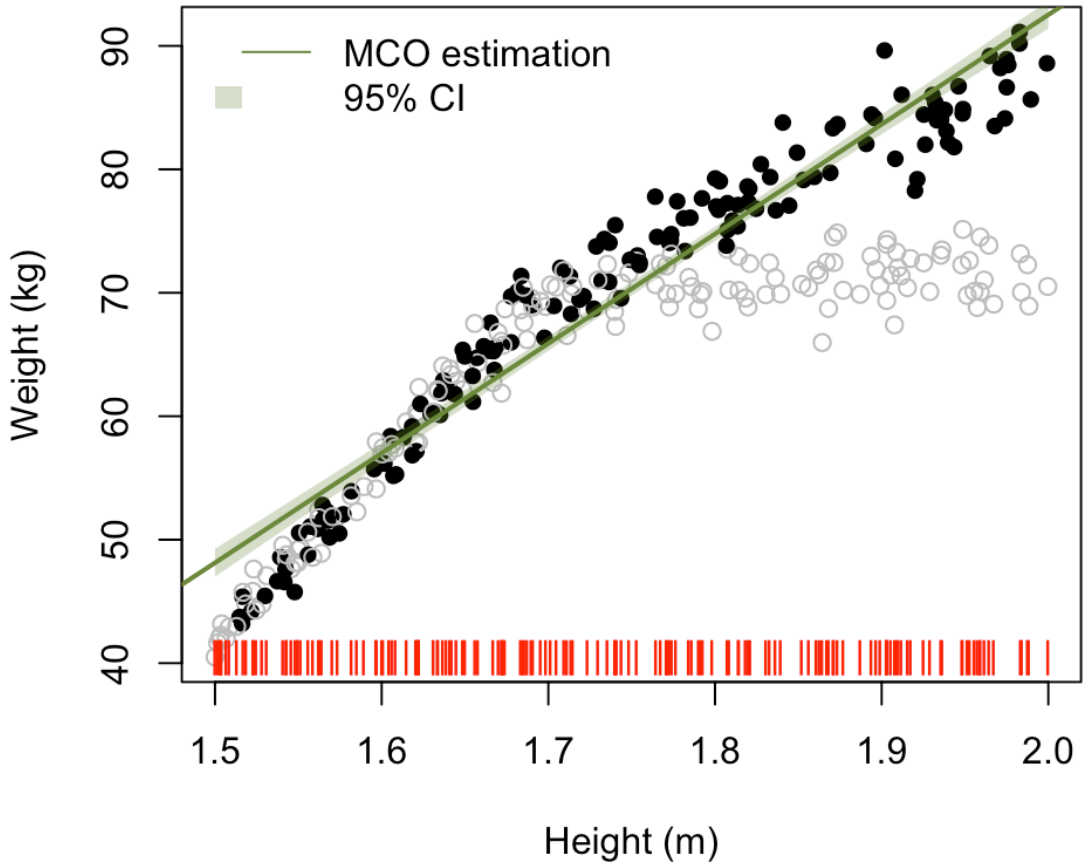
… pas plus de 20 pesées…



… saturation à 70kg…



… certains ne répondent pas…



Attention, une sous-population semble ne pas répondre…

… il n'y a aucun moyen de voir qu'il y a une structure de données manquantes ici!

En résumé, l'analyse cas complet n'est pas conseillée

- Peut introduire un biais dans l'estimation des paramètres,
- Peut réduire la puissance des tests statistiques en augmentant les écart-types des estimateurs,
- …

Solutions restantes (1)

Il reste de remplir les cases, c'est l'**imputation**.

- Estimer les **NA** avec des valeurs fixées (**Imputation simple**) :
  - Conditionnellement à la seule variable considérée :
    - **moyenne/médiane**,
    - **Last Observation Carried Forward** (**LOCF**),
  - Sur l'ensemble des variables :
    - k-plus proches voisins (**kNN** via),
    - forêts aléatoires (**missForest** (Stekhoven and Buehlmann 2012), itératif),
    - PCA (**missMDA** (Josse and Husson 2016), itératif),
    - …

Solutions restantes (2)

- **Imputation multiple** :
  - Modèles conditionnels (**mice** (van Buuren and Groothuis-Oudshoorn 2011))
  - PCA (**missMDA**)
  - …

Voir le CRAN Task View assez complet [1]

1 https://cran.r-project.org/web/views/MissingData.html

Mécanismes

Une classification des processus de perte de données

Due à Little and Rubin (1976) :

- **MCAR** (*Missing Completely At Random*), si la probabilité que la donnée soit manquante ne dépend pas de cette valeur ni des données observées

- **MAR** (*Missing At Random*), si la probabilité que la donnée soit manquante ne dépend pas de cette valeur, conditionellement aux données observées

- **MNAR** (*Missing Not At Random*), si la probabilité que la donnée soit manquante dépend de cette valeur, conditionellement aux données observées

Exemples

- **MCAR** Un capteur qui s'éteind et se rallume sans raison. Un patient malade qui se rend à l'hopital seulement lorsqu'il reçoit du courrier.

- **MAR** Un étudiant ne se rend pas à un examen de rattrapage puisqu'il a réussi l'examen normal.

- **MNAR** Qui gagne bien sa vie répond moins facilement à toute question sur son salaire.

Autre façon de le dire (1)

On qualifie le mécanisme dee perte de données en fonction de la probabilité de la perte de données conditionnellement aux données observées ou non observées :

- **MCAR** si la probabilité de manquement *est indépendante* des données manquantes et observées,

- **MAR** si la probabilité de manquement *est indépendante* de la donnée manquante, conditionnellement aux données observées,

- **MNAR** si la probabilité de manquement *n'est pas indépendante* de la donnée manquante, conditionnellement aux données observées.

Autre façon de le dire (2)

Si on note $\boldsymbol{M}$ la matrice indicatrice des données manquantes telle que

$$m_{i,j} = \begin{cases} 1 & \text{si la donnée } x_{i,j} \text{ est manquante} \\ 0 & \text{sinon} \end{cases}$$

pour l'individu i et la variable j.

Avec aussi, $\boldsymbol{X}^{(o)}$ les données observées et $\boldsymbol{X}^{(m)}$ les données manquantes, alors :

- Si $\mathbb{P}(\boldsymbol{M}|\boldsymbol{X}^{(o)}, \boldsymbol{X}^{(m)}) = \mathbb{P}(\boldsymbol{M})$, alors c'est un mécanisme **MCAR**,
- Si $\mathbb{P}(\boldsymbol{M}|\boldsymbol{X}^{(o)}, \boldsymbol{X}^{(m)}) = \mathbb{P}(\boldsymbol{M}|\boldsymbol{X}^{(o)})$, alors c'est un mécanisme **MAR**,
- Si $\mathbb{P}(\boldsymbol{M}|\boldsymbol{X}^{(o)}, \boldsymbol{X}^{(m)}) \neq \mathbb{P}(\boldsymbol{M}|\boldsymbol{X}^{(o)})$, alors c'est un mécanisme **MNAR**.

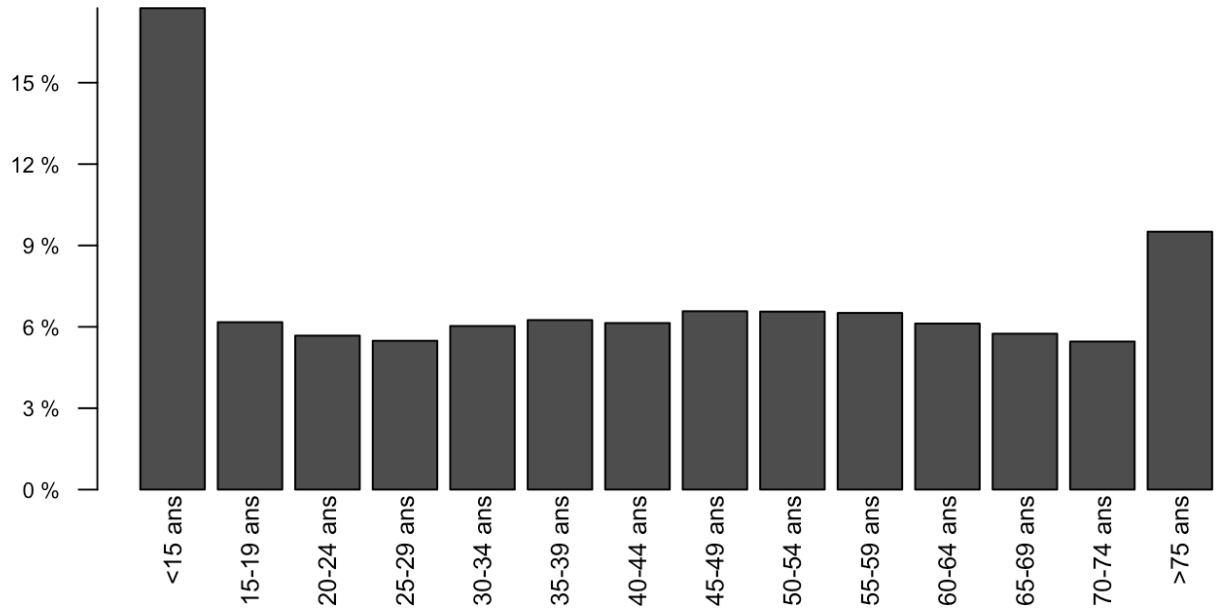
Commentaires

- **MCAR** Le plus “simple” à gérer, mais peu réaliste
- **MAR** Peut aussi être géré si approche multivariée
- **MNAR** On s’en rend compte après coup car rien ne permet de le savoir dans les données

Exemple, l'âge de la population française selon l'Insee (1)

On peut télécharger ces données sur le site de l'Insee[1].

▶ Code



1 https://www.insee.fr/fr/statistiques/2381474#figure1_radio1

Exemple, l'âge de la population française selon l'Insee (2)

Exercice

Vous créez un jeu de données de taille $n = 1000$ formé par deux variables :

- age (variable continue positive) à partir de la distribution de l'Insee,
- a_moins_de_15_ans (1 si oui, 0 sinon) à partir de la variable age

A partir de ce jeu de données, vous allez simuler trois scénarios de données manquantes :

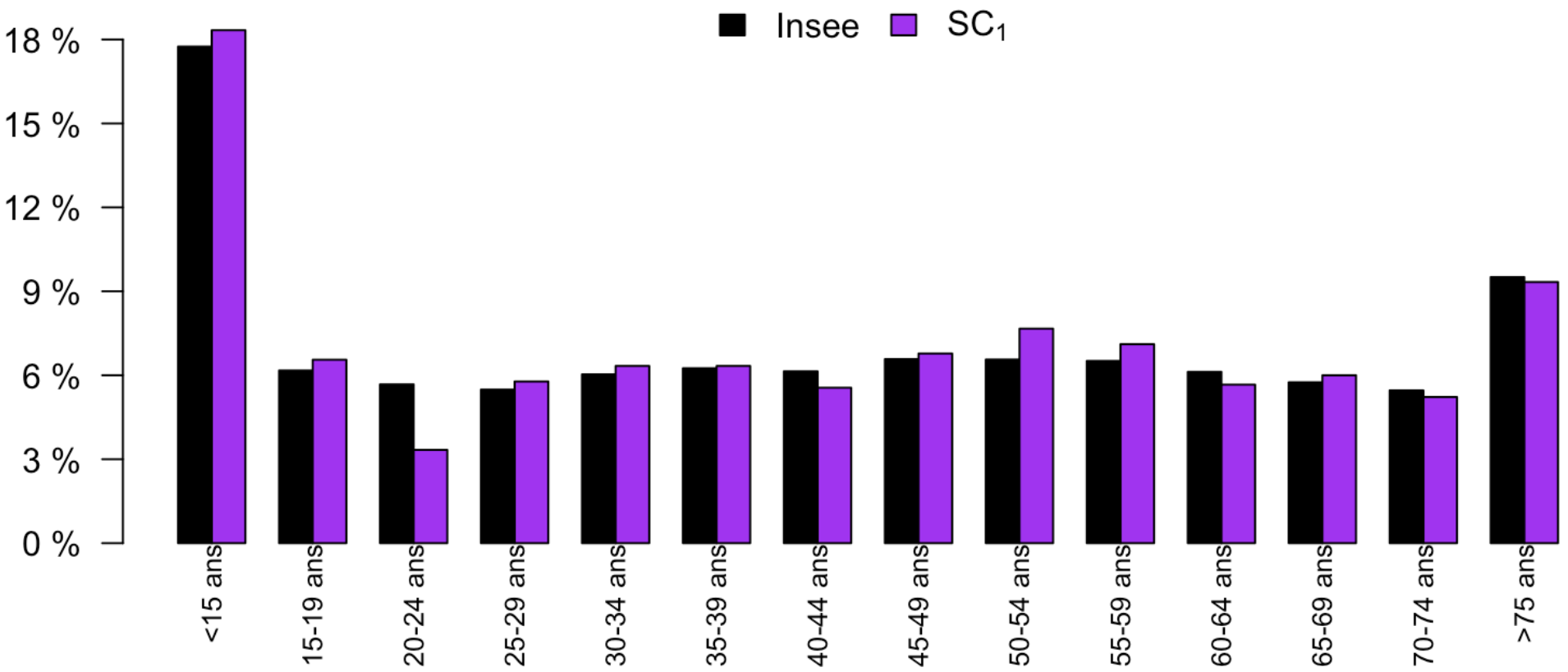
- Sc_1 : "10% des gens refusent de donner leur âge."
- Sc_2 : "Un bug informatique a supprimé les âges des moins de 15 ans."
- Sc_3 : "C'est vendredi, les étudiants ne sont pas là!"

Vous discuterez des hypothèses prises et des résultats obtenus.

Sc_1 : “10% des gens refusent de donner leur âge.”

10% des observations sont à retirer du jeu de données

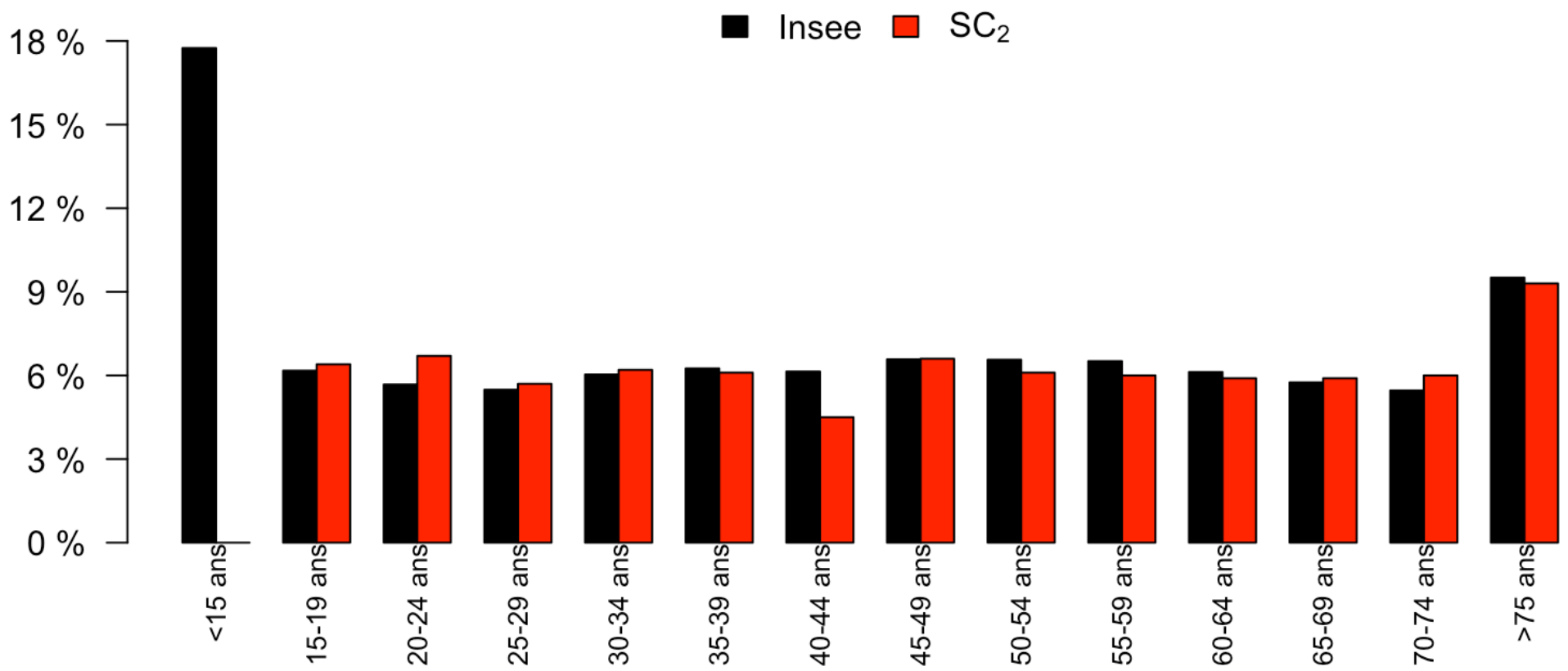
▶ Code



Sc_2 : “Suppression des âges des moins de 15 ans.”

Les individus de classe 1 sont à retirer du jeu de données (hyp 1)

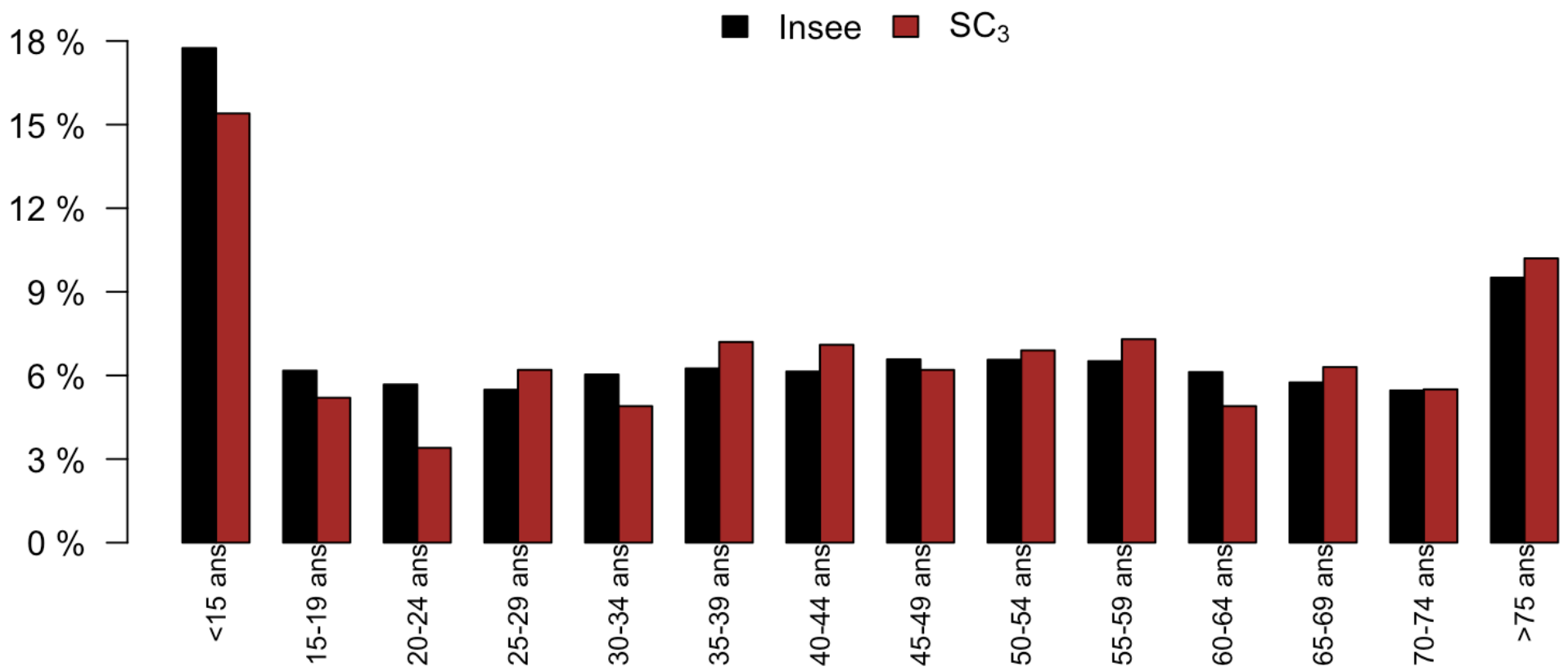
▶ Code



Sc_3 : “C’est vendredi, les étudiants ne sont pas là!”

La moitié des individus de classe 3 est à retirer du jeu de données (hyp 2)

▶ Code



Exemple, l'âge de la population française selon l'Insee, commentaires

- Sc_1 : On reconnaît l'hypothèse **MCAR**
- Sc_2 peut être géré conditionellement à la variable indicatrice : **MAR**
- Sc_3 est plus difficile à gérer : **MNAR**

Une solution pour le cas MNAR ?

Rechercher des informations supplémentaires via une autre variable par exemple ou un autre essai similaire ou l'expérience d'un expert

Imputation ou prédiction ?

Modèle linéaire simple à une covariable

On considère le modèle suivant :

$$\left[y|x,\alpha,\beta,\sigma^2\right] \sim \mathcal{N}(\alpha+\beta x,\sigma^2),$$

où x est la covariable, α l'ordonnée à l'origine, β la pente et σ^2 la variance du bruit additif. C'est le modèle linéaire très classique.

Ici $\theta = (\alpha,\beta,\sigma^2)$ et des données manquantes apparaîssent dans y.

La problématique ressemble à un problème d'estimation de paramètres dans un modèle linéaire simple, mais avec des données manquantes.

Simulation des données

Soit le modèle de simulation

$$y = x + \epsilon$$

où $x \sim \mathcal{N}(0, 1), \epsilon \sim \mathcal{N}(0, 1/4)$ et $\epsilon \perp\!\!\!\perp x$. Donc $\alpha = 0, \beta = 1$ et $\sigma^2 = 1/4$. On a accès à :

- un échantillon d'entraînement $S = (x_i, y_i)_{i=1,\dots,n}$ de taille $n = 50$,
- un échantillon de test $\tilde{S} = (x_i)_{i=n+1,\dots,n+\tilde{n}}$ de taille $\tilde{n} = 30$.

En régression, on réalise deux opérations successives :

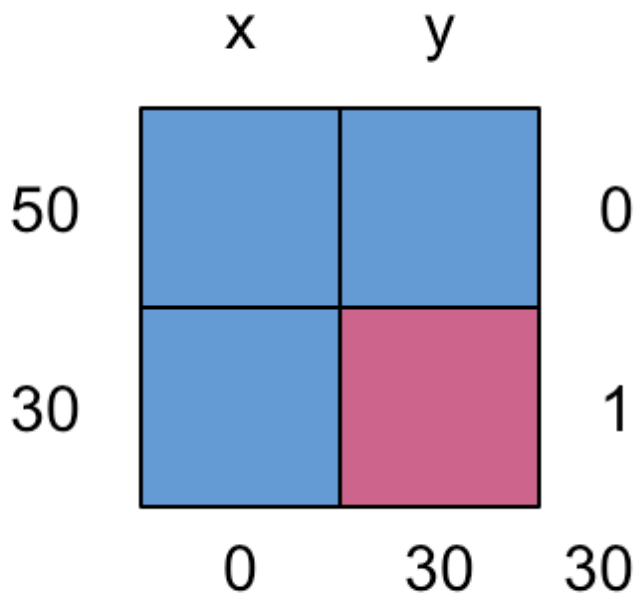
- construction d'un modèle de régression sur S, noté $\mathcal{P}$,
- estimation de la réponse de $\mathcal{P}$ à une covariable x_0, notée $\hat{y_0} = \mathcal{P}(x_0)$.

Lien avec les données manquantes

En combinant S et $\tilde{S}$, on a accès à un jeu de données de dimensions $(n + \tilde{n}) \times (2)$ avec $\tilde{n}$ données manquantes : les valeurs $(y_i)_{i=n+1,\ldots,n+\tilde{n}}$.

On peut observer la structure des données manquantes grâce aux commandes suivantes

▶ Code



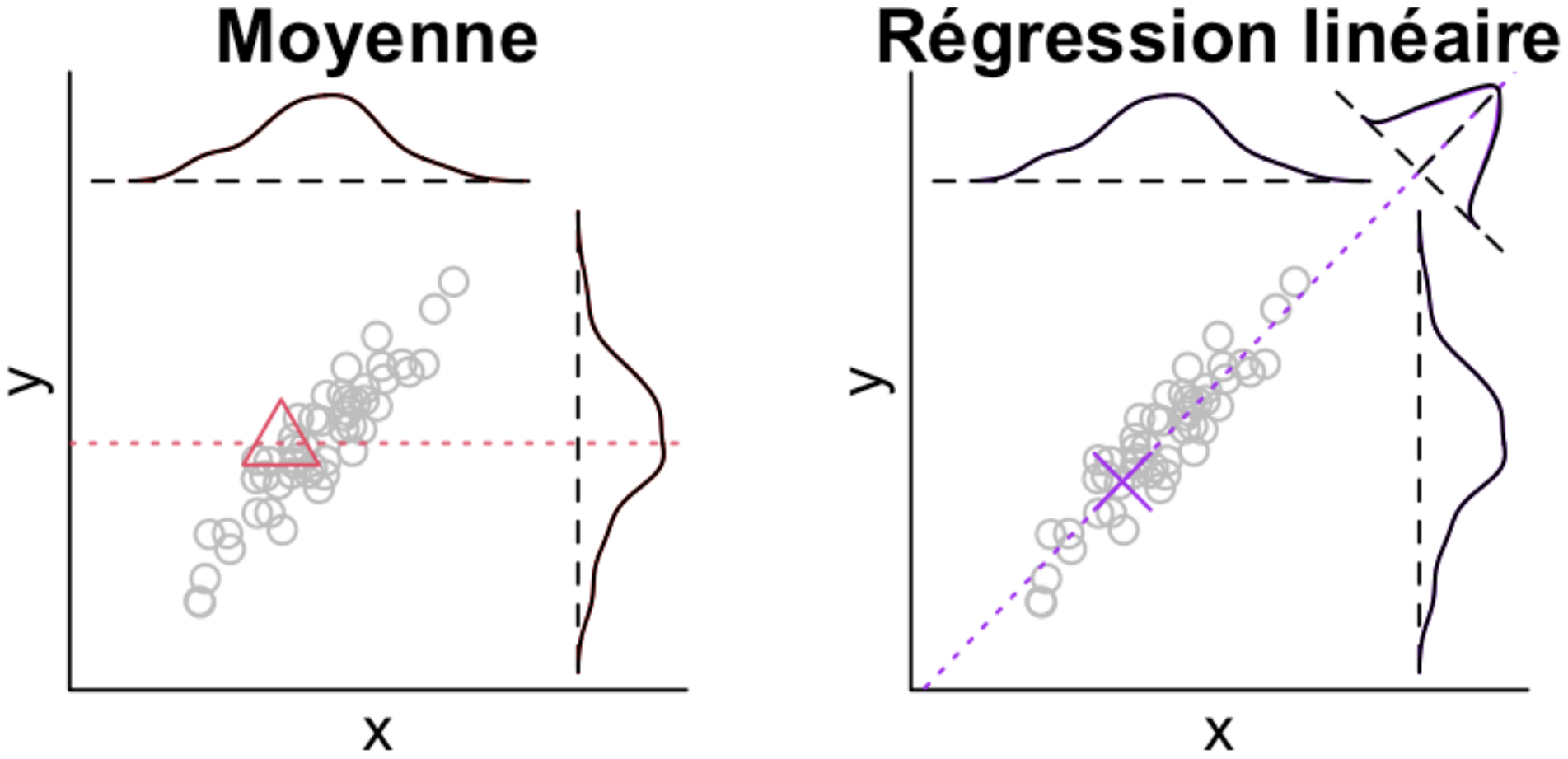
Analyse du jeu de données simulées

Comparaison de 2 méthodologies d'imputation :

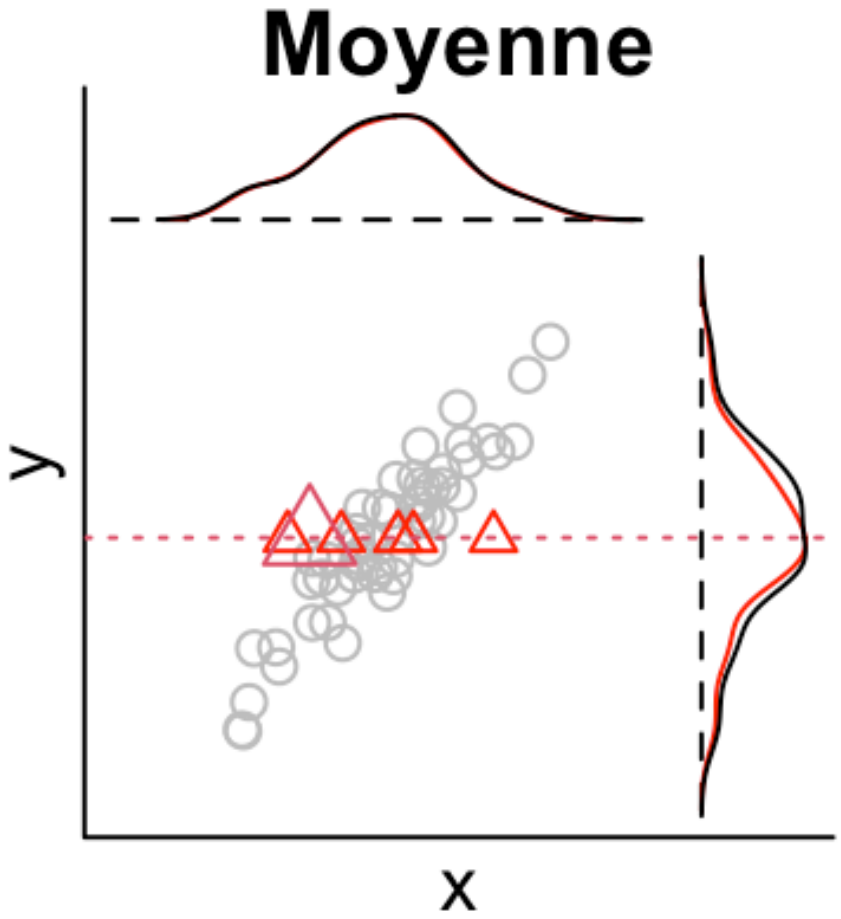
- à la moyenne,
- en utilisant le modèle de régression linéaire.

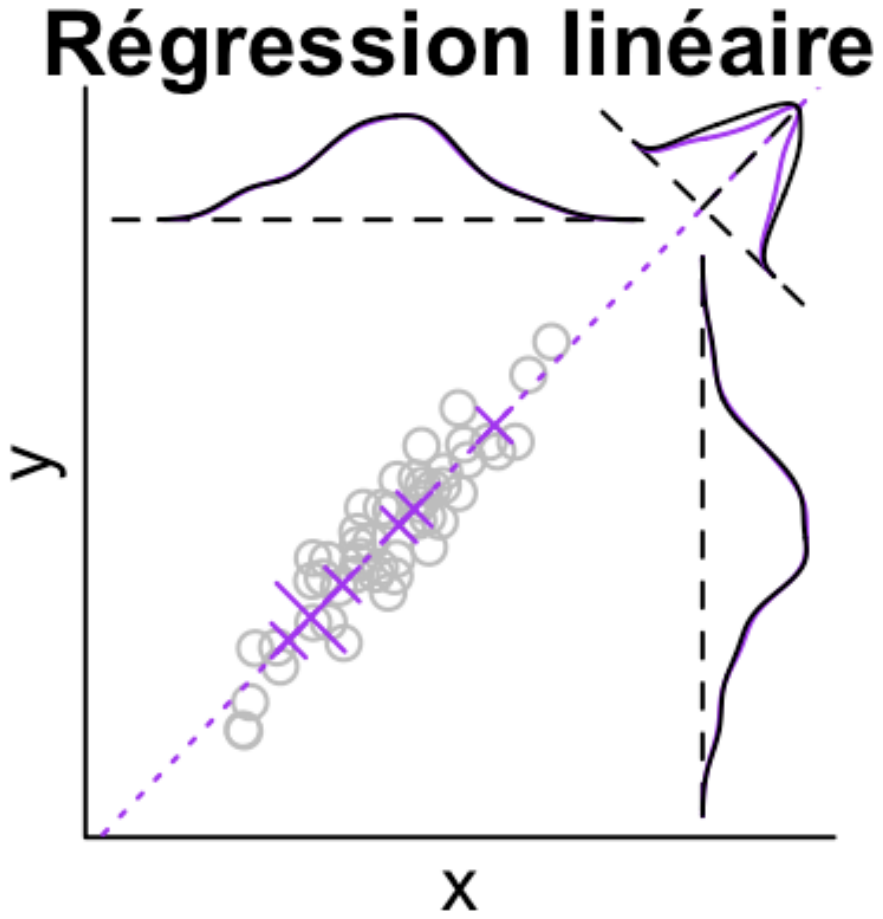
Dans chaque cas on observera attentivement le comportement des distributions après imputations.

Après imputation d'une observation

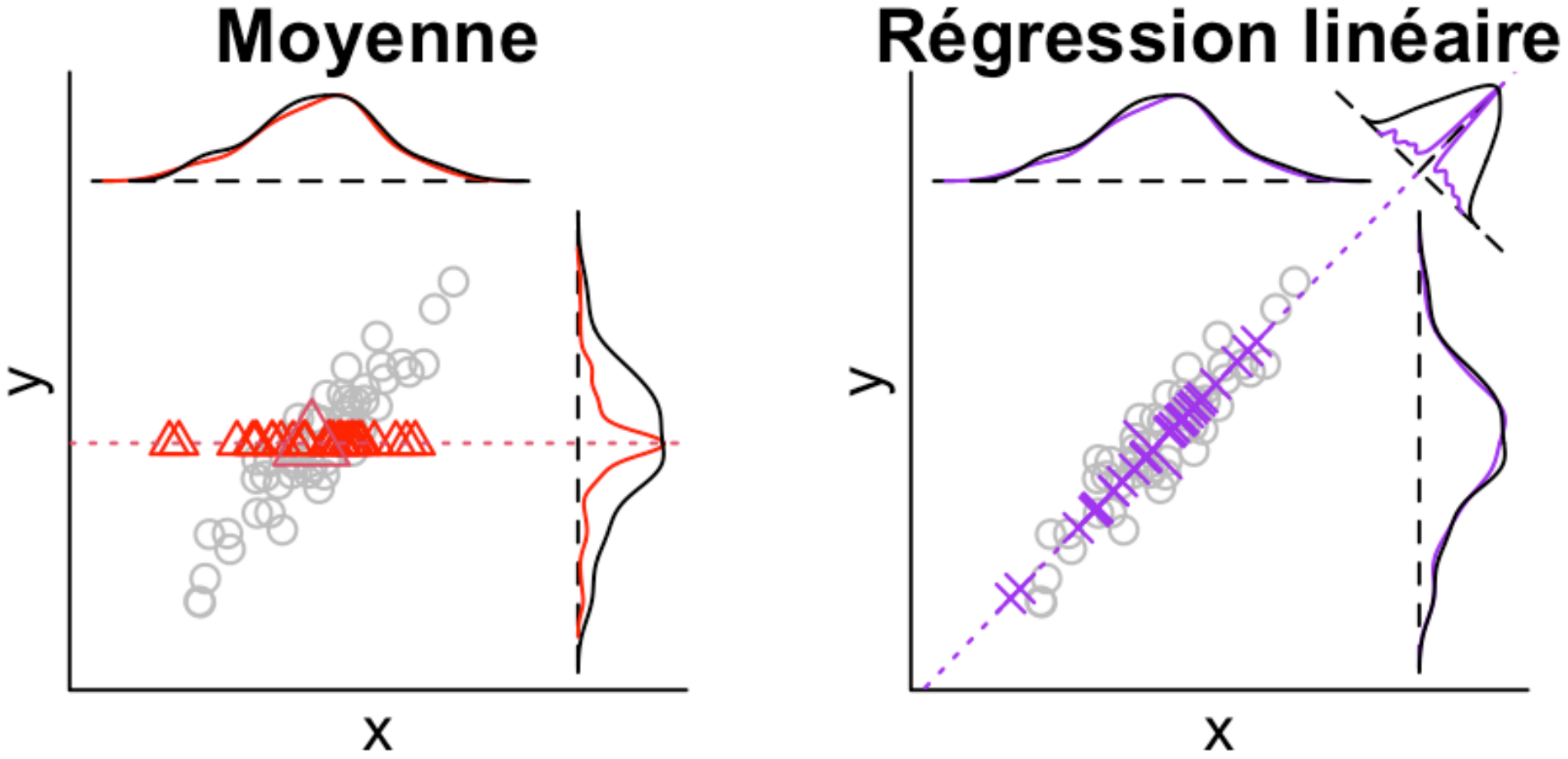


Après imputation de 5 observations

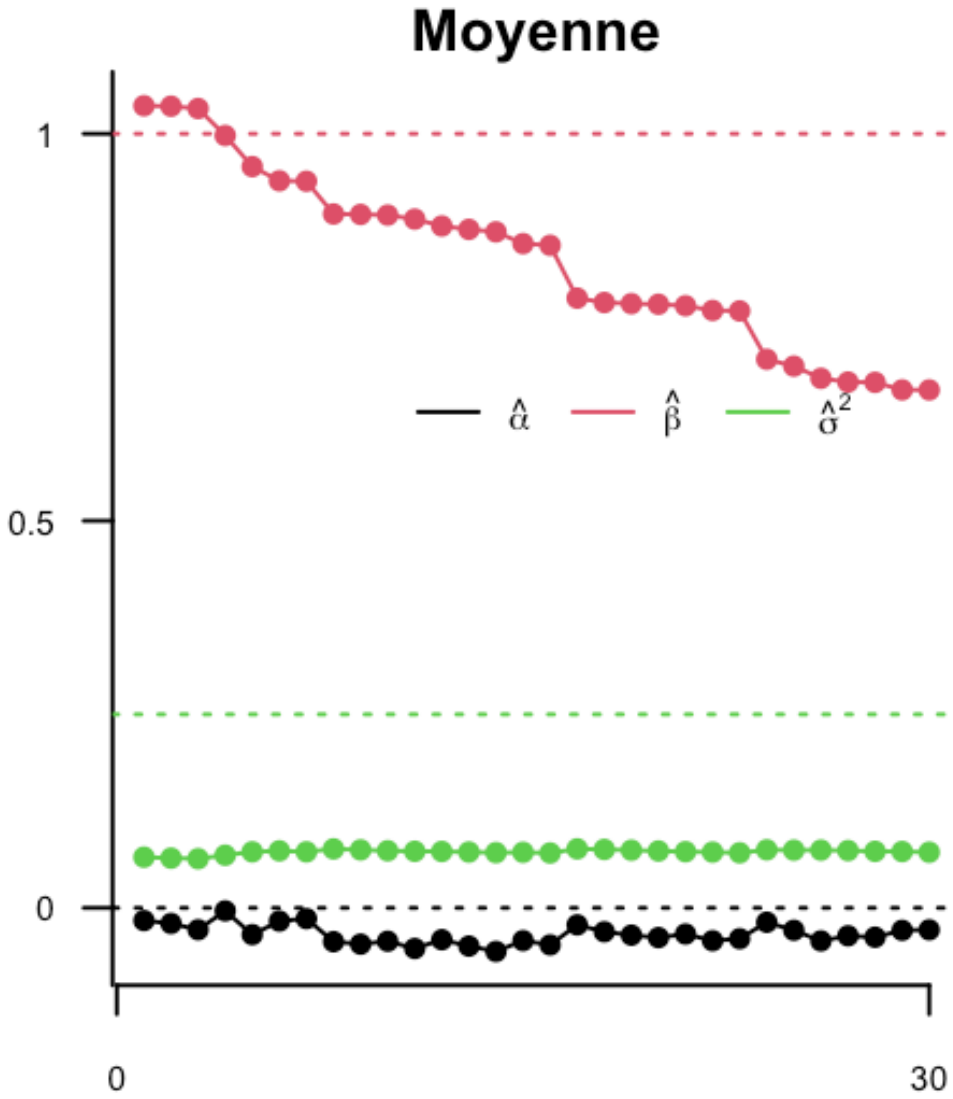


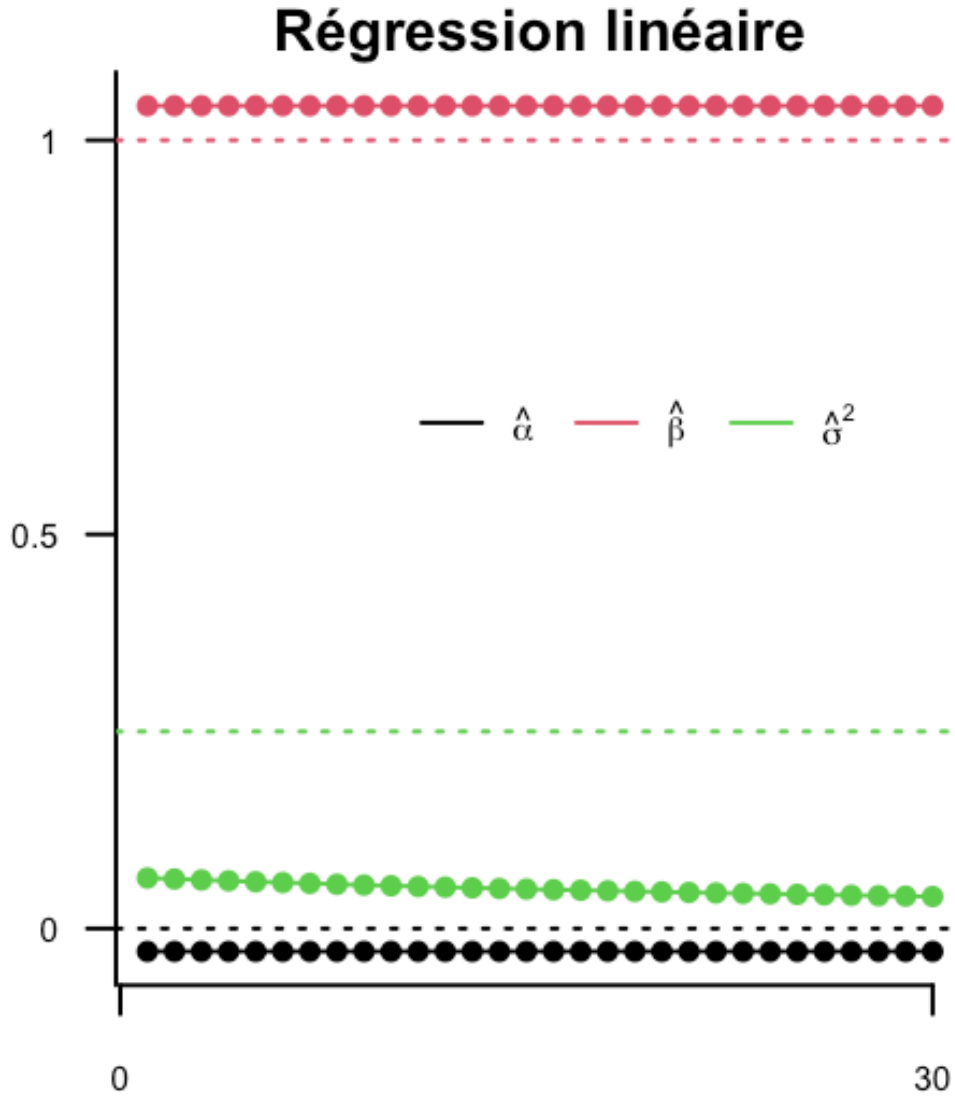


Après imputation de 30 observations



Au total


Moyenne
1
0.5
0
0
30
$\hat{\alpha}$
$\hat{\beta}$
$\hat{\sigma}^2$


Régression linéaire
1
0.5
0
0
30
$\hat{\alpha}$
$\hat{\beta}$
$\hat{\sigma}^2$

Observations

Deux impératifs apparaissent :

- Conditionner l'estimation des données manquantes sur les données observées.
- Utiliser un modèle de régression adapté.
- Garder en tête que l'ensemble reconstruit doit être réutilisé : c'est la grosse différence avec la régression/classification/analyse classique.

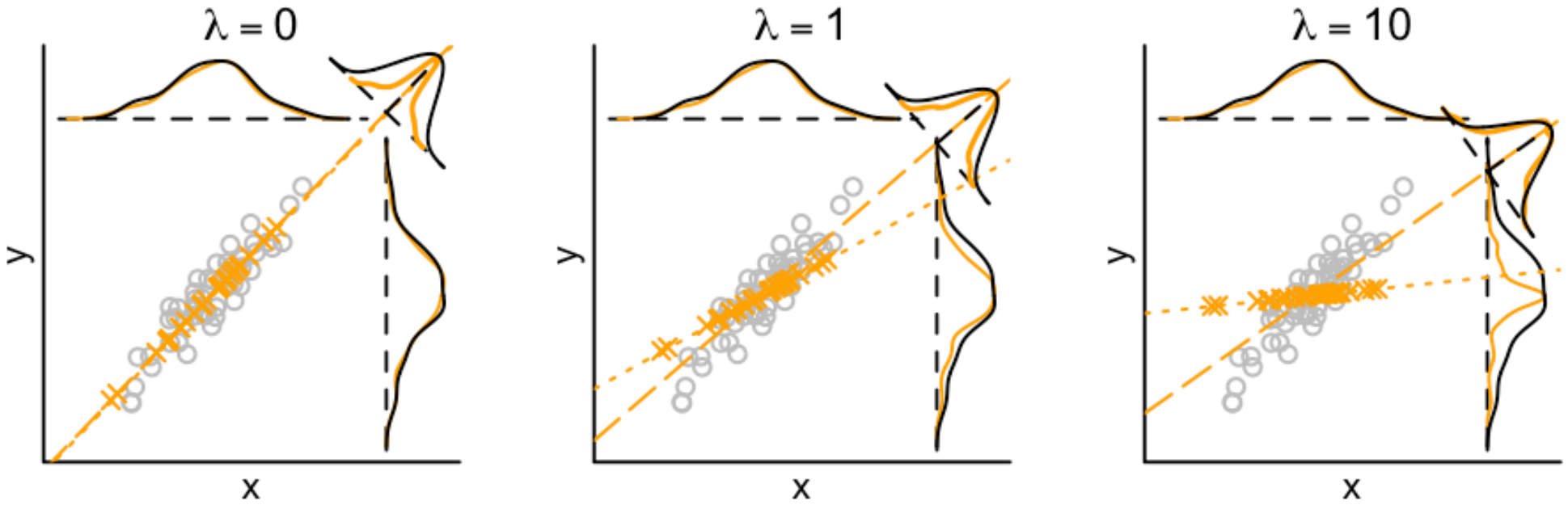
Une solution ?

- Pourquoi ne pas faire de la régularisation ?

$$\text{Régularisation Ridge : } \min_{\alpha,\beta} (y - \alpha - x\beta)^2 + \lambda\beta^2$$

La régularisation

- Régression linéaire pénalisée Ridge pour l'imputation.
- Régression linéaire pour l'estimation des paramètres.



⟶ Compromis biais-variance.

La régularisation (2)

Conclusion

La régularisation permet de réduire l'erreur que l'on fait sur l'estimation de σ^2 au détriment de la variance estimée de y

La prédiction, qui est une espérance conditionelle, ne prend pas en compte l'incertitude liée au modèle (bruit d'observation,…)

Une solution ?

Il faut utiliser une méthode qui introduise cette incertitude

Générer M jeux de données imputés pour obtenir cette variabilité

C'est **la régression stochastique**

La régression stochastique - *improper imputation*

Alors que le modèle de prédiction était précédemment

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

où $\hat{\beta}$ était estimé sur le jeu de données S, nous allons maintenant utiliser l'estimateur

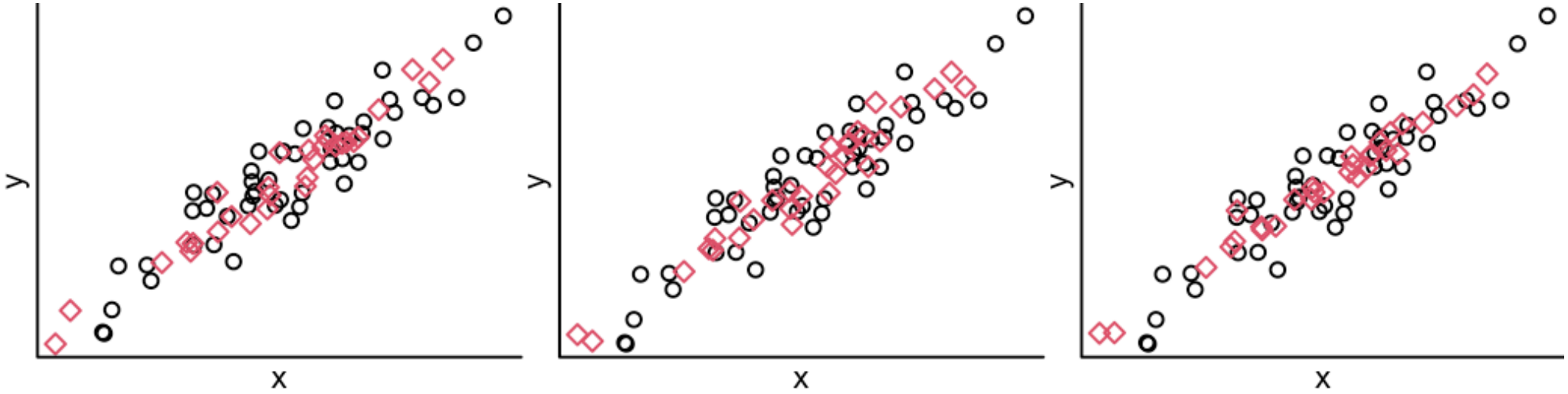
$$\tilde{y} = \hat{\alpha} + \hat{\beta}x + \eta,$$

avec $\eta \sim \mathcal{N}(0, \hat{\sigma}^2)$ et $\hat{\sigma}^2$ est estimé sur S via

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

La régression stochastique - *improper imputation* (2)

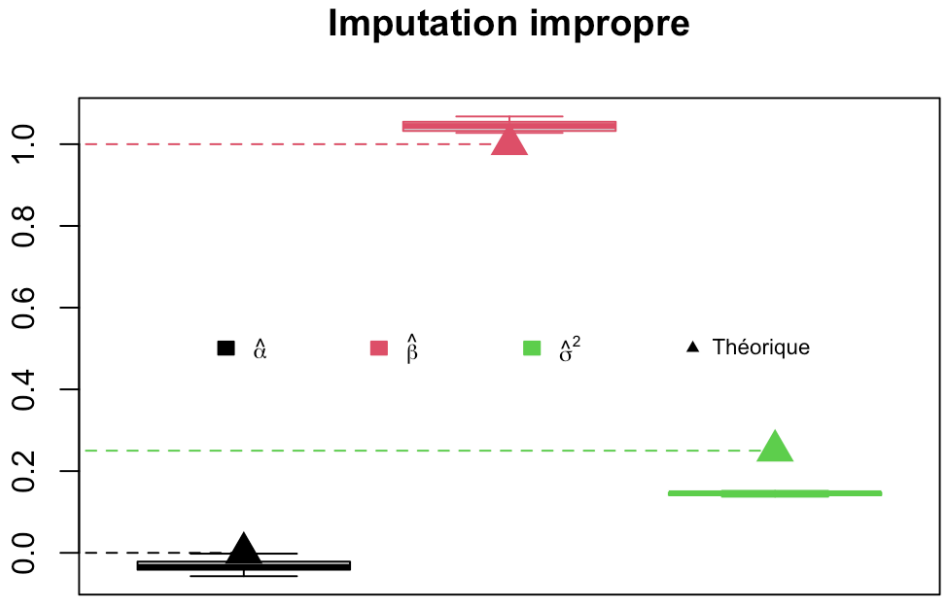
On répète un nombre $M = 20$ d'imputations du jeu de données intitiales. En voici 3 versions.



La moyenne empirique des $M = 20$ estimations pour chacun des 3 coefficients:

$$\hat{\alpha}_N \approx -0.032 \pm 0.043,\ \hat{\beta}_N \approx 1.045 \pm 0.041,\ \hat{\sigma}^2_N \approx 0.145$$

Le boxplot, toujours aussi utile



Les points théoriques **sont**[1] en dehors des distributions… A-t-on oublié quelque chose ?

1 Très probablement

L'imputation multiple - *proper* (1)

Problème

Cette solution ne prend pas en compte la variabilité sur les paramètres.

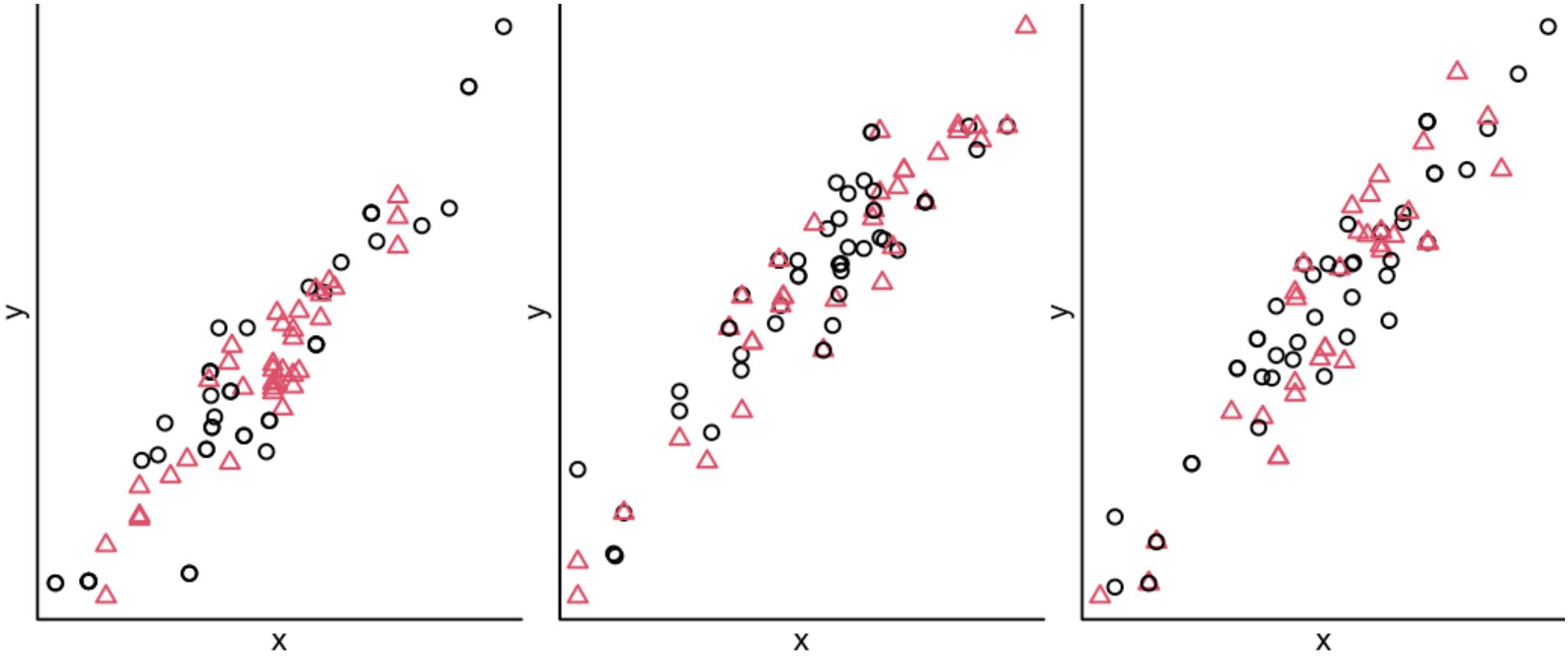
Deux solutions

- **Bayes** : Tirer M paramètres sur la distribution a posteriori (si définie…). Estimer M jeux de données différents en suivant ces paramètres Tanner and Wong (1987), utilisée dans **MICE** van Buuren and Groothuis-Oudshoorn (2011)
- **Bootstrapper** $(S, \tilde{S})$ pour créer M jeux de données dans lesquels il y a potentiellement des données manquantes, approche de **missMDA** Josse and Husson (2016).

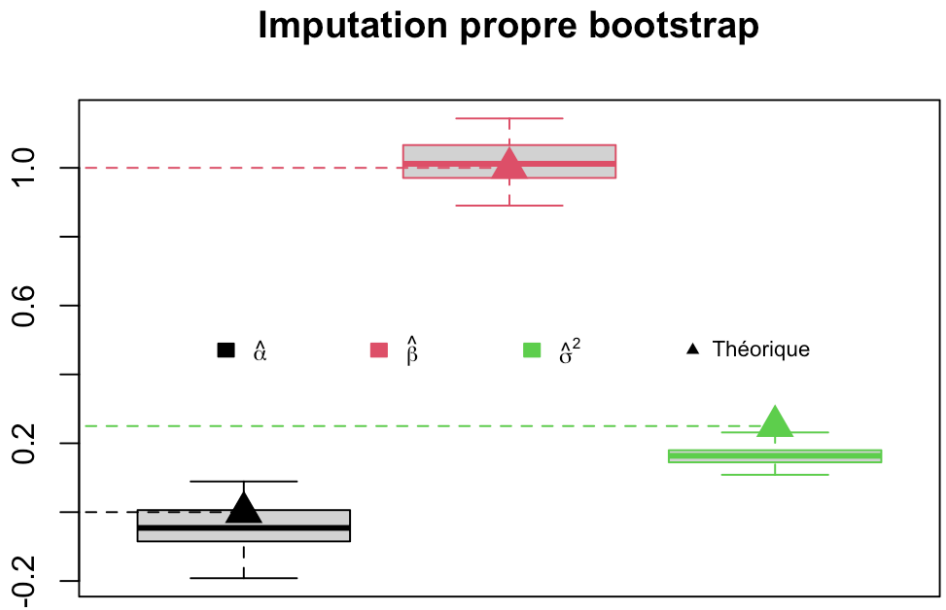
- Ensuite appliquer la méthode d'analyse sur chacun des jeux de données séparément.
- Estimer les paramètres par aggrégation des M modèles.

L'imputation multiple - *proper* (2)

3 jeux de données complétés pour l'approche par bootstrap



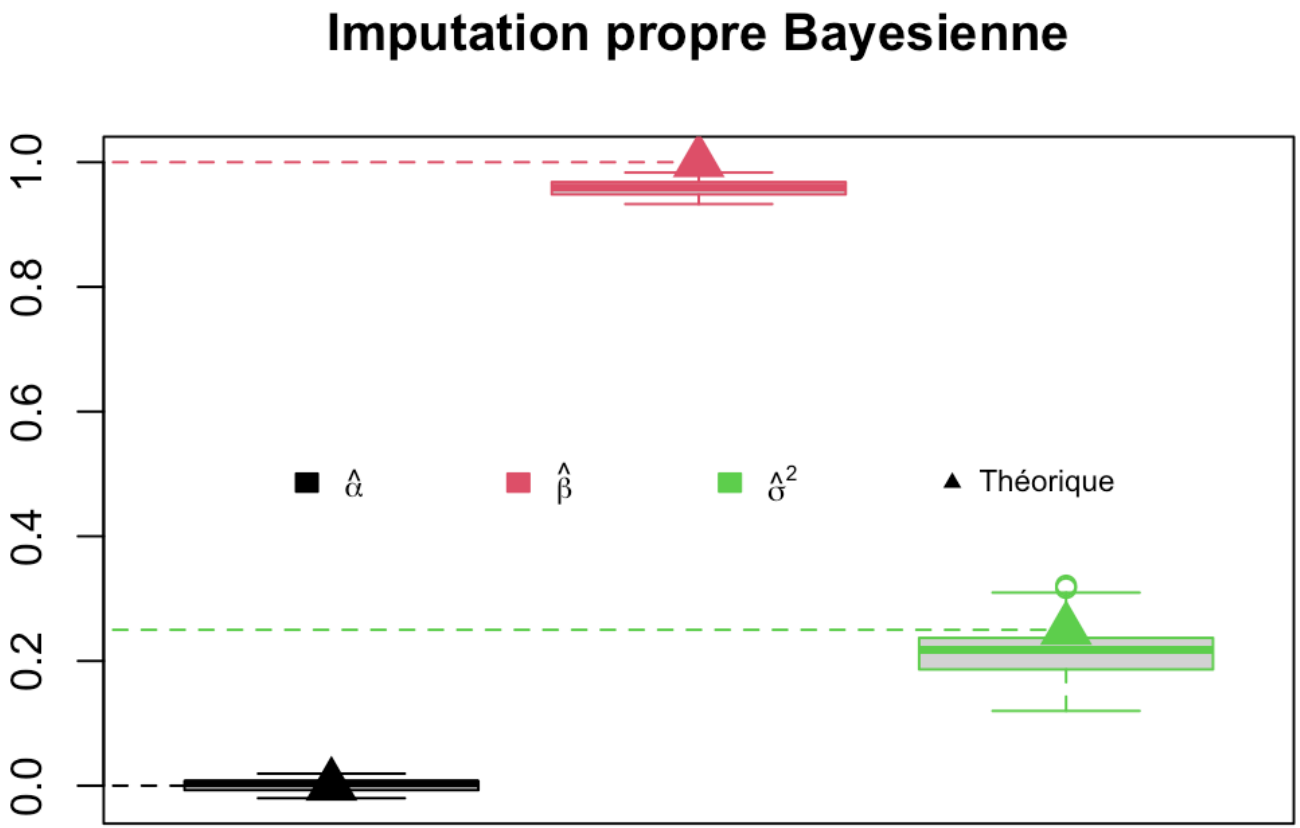
L'imputation multiple - *proper* (3)



Les points théoriques **ne sont pas**[1] en dehors des distributions… on limite le biais

1 Très probablement

L'imputation multiple - *proper* (4)



Biais là aussi limité

L'imputation multiple - *proper* (5)

On obtient des estimations :

- Via bootstrap :

$$\hat{\alpha}_{\text{Boot}} \approx -0.039 \pm 0.045,\ \hat{\beta}_{\text{Boot}} \approx 1.019 \pm 0.045,\ \hat{\sigma}^2_{\text{Boot}} \approx 0.163$$

- Via la postérieure/bayésienne :

$$\hat{\alpha}_{\text{Bayes}} \approx 0.001 \pm 0.01,\ \hat{\beta}_{\text{Bayes}} \approx 0.958 \pm 0.013,\ \hat{\sigma}^2_{\text{Bayes}} \approx 0.216$$

Pour rappel, dans les cas impropres :

Moyenne	$\hat{\alpha}_0 \approx -0.028 \pm 0.072$	$\hat{\beta}_0 \approx 0.669 \pm 0.069$	$\hat{\sigma}^2_0 \approx 0.411$
Régression linéaire	$\hat{\alpha}_l \approx -0.029 \pm 0.041$	$\hat{\beta}_l \approx 1.044 \pm 0.039$	$\hat{\sigma}^2_l \approx 0.131$
Régression stochastique	$\hat{\alpha}_N \approx -0.032 \pm 0.043$	$\hat{\beta}_N \approx 1.045 \pm 0.041$	$\hat{\sigma}^2_N \approx 0.145$

Règles de Rubin (*Rubin's Rules*)

Soit M jeux de données imputés alors la règle de Rubin stipule que

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m,$$

où $\hat{\theta}_m$ est la valeur du paramètre estimé pour le jeu de donné d'indice m. Il vient alors que la variance de l'estimateur $\hat{\theta}$, selon les règles de Rubin, est :

$$\widehat{\mathbb{V}}^2_{Tot} = W + (1 + \tfrac{1}{M})B,$$

$$\text{Variance imputée } (\textit{within}) \text{: } W = \tfrac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{SE}}^2,$$

Règles de Rubin dans le cas du modèle linéaire (1)

On note $(\hat{\alpha}_m, \hat{\beta}_m, \hat{\sigma}^2_m)$ l'estimateur du paramètre $(\alpha, \beta, \sigma^2)$ sur le jeu de données m et $\mathrm{SE}_m = (\mathrm{SE}_{\hat{\alpha}_m}, \mathrm{SE}_{\hat{\beta}_m}, \mathrm{SE}_{\hat{\sigma}^2_m})$ l'erreur standard associée aux paramètres. On a donc :

$$(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2) = \left(\frac{1}{M} \sum_{m=1}^{M} \hat{\alpha}_m, \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m, \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^2_m \right),$$

et la variance de l'estimateur $\hat{\theta}_{Tot}$, selon les règles de Rubin, est :

$$\widehat{\mathbb{V}}^2_{Tot,j} = \widehat{W}_j + (1 + 1/M)\widehat{B}_j,$$

$$\text{Variance imputée } (\textit{within}) \text{: } \widehat{W}_j = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{SE}}(\hat{\theta}_{j,m})^2,$$

$$\text{Variance inter-imputation } (\textit{between}) \text{: } \widehat{B}_j = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_{j,m} - \hat{\theta}_j)^2,$$

pour $j = \alpha, \beta, \sigma^2$.

Règles de Rubin dans le cas du modèle linéaire (2)

Dans le modèle linéaire on sait, cf **M1 modèle linéaire** :

Statistique	Estimateur
$\mathrm{SE}(\hat{\beta}_m)^2 = \frac{\sigma_m^2}{\sum_{i=1}^n (x_{i,m} - \bar{x}_m)^2}$	$\widehat{\mathrm{SE}}(\hat{\beta}_m)^2 = \frac{\hat{\sigma}_m^2}{\sum_{i=1}^n (x_{i,m} - \bar{x}_m)^2},$
$\mathrm{SE}(\hat{\alpha}_m)^2 = \sigma_m^2 \left(\frac{1}{n} + \frac{\bar{x}_m^2}{\sum_{i=1}^n (x_{i,m} - \bar{x}_m)^2} \right)$	$\widehat{\mathrm{SE}}(\hat{\alpha}_m)^2 = \hat{\sigma}_m^2 \left(\frac{1}{n} + \frac{\bar{x}_m^2}{\sum_{i=1}^n (x_{i,m} - \bar{x}_m)^2} \right),$
$\mathrm{SE}(\hat{\sigma}_m^2)^2 = \frac{2\sigma_m^4}{n-2}$	$\widehat{\mathrm{SE}}(\hat{\sigma}_m^2)^2 = \frac{2\hat{\sigma}_m^4}{n-2}.$

Paquetages et implémentations

Un passage à l'échelle nécessaire

Nous avons vu les principales caractéristiques de l'imputation par comparaison à la régression classique.

En pratique, les jeux de données ne sont pas bivariés et les NA ne sont pas que dans une seule variable…

Dommage…

Des modifications de ce que nous venons de voir ont été imaginées afin de gérer la présence de NA dans le cas général.

missMDA

- Utilisation de méthodes factorielles.
- Variables qualitatives/quantitatives/mixtes.
- Imputation simple/multiple.
- Echantillonnage bootstrap/bayésienne.
- Beaucoup de choses à dire…

Amelia II (algorithme EMB), Honaker and King (2010)

- Hypothèse d'un jeu de donnée complété qui suit une distribution normale multivariée (p variables) de paramètre (μ, Σ).
- Bootstrap ou bayésienne

Echantillonnage bootstrap et imputation via algorithme EM… EMB(ootstrap).

mice

Modèles conditionnels chaînés.

Utilise une pénalisation Ridge dans l'imputation, paramètre κ de l'ordre de $\kappa = 0.0001$ ($\kappa = 0.1$ est grand dans ce cas et *"may introduce a systematic bias toward the null, and should thus be avoided"*)

4 modèles d'imputation :

- `norm.predict` : imputation sans alea,
- `norm.nob` : imputation stochastique,
- `norm` : imputation stochastique et postérieure/bayésienne,
- `norm.boot` : Estimtation des paramètres sur un échantillon bootstrap des données observées.

mice (2)

Les variables sont organisées par nombre d'obs. NA croissant. Un modèle sur la première variable conditionellement aux autres est construit. Les données manquantes sont imputées grâce à ce modèle. C'est au tour de la variable suivante etc… On recommence jusqu'à convergence.

Voir Vignette[1].

1 https://stefvanbuuren.name/fimd/sec-linearnormal.html#def:norm

missForest

- Utilisation de “random forests”
- Variables qualitatives/quantitatives/mixtes.
- Imputation simple.

Les variables sont organisées par nombre d’obs. NA croissant. Un modèle sur la première variable conditionellement aux autres est construit, sur les observations présentes pour cette variable. Les données manquantes sont imputées grâce à ce modèle. C’est au tour de la variable suivante etc… On recommence jusqu’à convergence.

k-NN, Troyanskaya et al. (2001)

- Non-itérative.
- Imputation simple. Pour une observation avec des NA, utilise les observations La fonction `impute.knn` du package `impute`. Le paramètre le plus important est `k`, le nombre de plus proches voisins.
- Package `impute` : `kNN` réalisé sur les variables et non les observations. Données quantitatives. Distance euclidienne.
- Package `VIM`. Données mixtes. Distance de Gower.

Conclusion

Conclusion

C’est un sujet difficile et/car computationnel, lire :

- Imbert and Vialaneix (2018), une biblio très fournie méthodo et implémentations.
- Imputation multiple & analyse factorielle, François Husson[1].
- Utilisation d’Amelia II et imputation multiple[2]

1. http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/105543_museum_hist_nat.pdf
2. https://cran.r-project.org/web/packages/Amelia/vignettes/intro-mi.html

Références

Honaker, James, and Gary King. 2010. “What to Do about Missing Values in Time Series Cross-Section Data.” *American Journal of Political Science* 54 (3): 561–81.

Imbert, Alyssa, and Nathalie Vialaneix. 2018. “Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes.” *Journal de La Societe Française de Statistique* 159 (2): 1–55. https://hal.inrae.fr/hal-02618033.

Josse, Julie, and François Husson. 2016. “missMDA: A Package for Handling Missing Values in Multivariate Data Analysis.” *Journal of Statistical Software* 70 (1): 1–31. https://doi.org/10.18637/jss.v070.i01.

Little, Roderick JA, and Donald B Rubin. 1976. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

Stekhoven, Daniel J., and Peter Buehlmann. 2012. “MissForest - Non-Parametric Missing Value Imputation for Mixed-Type Data.” *Bioinformatics* 28 (1): 112–18.

Tanner, Martin A., and Wing Hung Wong. 1987. “The Calculation of Posterior Distributions by Data Augmentation.” *Journal of the American Statistical Association* 82 (398): 528–40.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. 2001. “Missing value estimation methods for DNA microarrays .” *Bioinformatics* 17 (6): 520–25.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate Imputation by Chained Equations in r.” *Journal of Statistical Software* 45 (3): 1–67. https://www.jstatsoft.org/v45/i03/.