Données Manquantes

que peut-on faire ?

Hadrien Lorenzo

DU BDSI 16-17/10/2021

Section 1

Introduction

Analyse cas complet

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),
- forte proportion de NA.

On peut aussi retirer les variables associées aux NA si :

- la structure des données le permet (p important)
- 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.

Solutions restantes

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 `FNN`),
    - forêts aléatoires (**missForest** (Stekhoven & Buehlmann, 2012), itératif),
    - SVD (**missMDA** (Josse & Husson, 2016), itératif),
    - ...
- **Imputation multiple** :
  - Modèles conditionnels (**mice** (van Buuren & Groothuis-Oudshoorn, 2011))
  - SVD (**missMDA**)
  - ...

Une classification des processus de perte de données

Due à Little & Rubin (1976).

MCAR (*Missing Completely At Random*), si la probabilité de perte de l'information est la même pour toutes les observations et donc ne dépend ni des données observées ni des autres données manquantes.

MAR (*Missing At Random*), si la probabilité que la donnée soit manquante est associée à une ou plusieurs autres variables, observées.

MNAR (*Missing Not At Random*), si la probabilité de ne pas observer cette valeur dépend de cette valeur et ou dépend de variables non 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 (dur à prédire/observer donc...)

MAR Un végétarien qui ne donne pas son avis (gustatif) sur un morceau de viande qui lui est présenté. Un nourisson qui ne renseigne pas la couleur du cheval blanc d'Henri IV.

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

Commentaires

MCAR Le plus "simple" à gérer, mais peu réaliste.

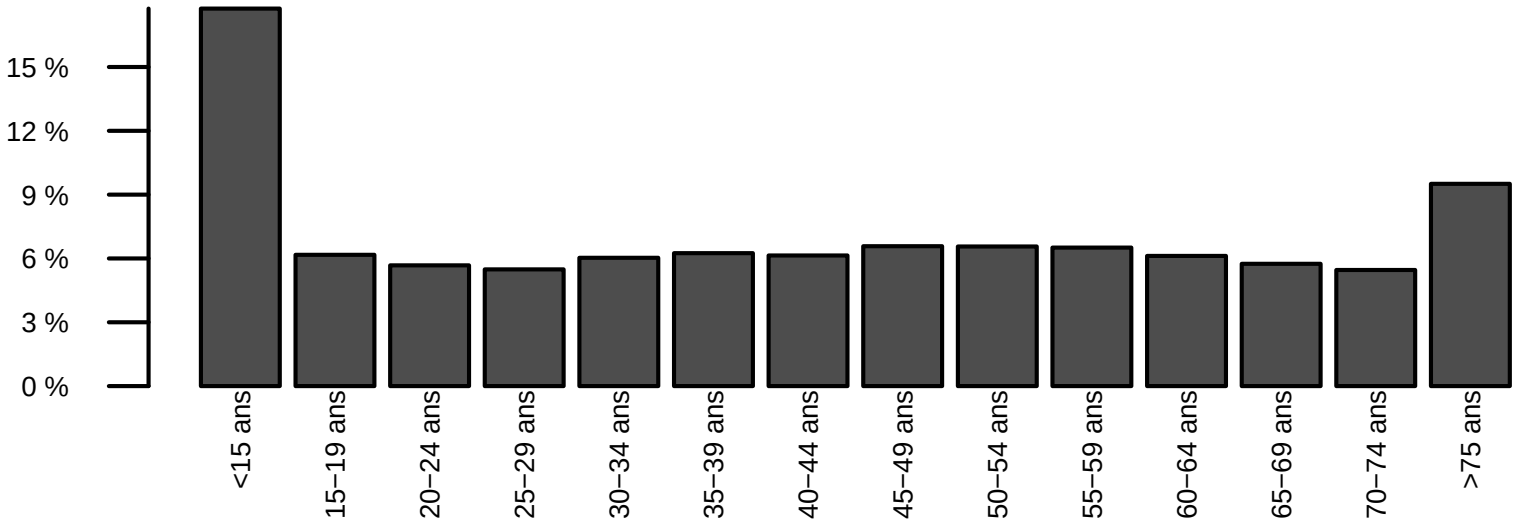
MAR Peut aussi être géré si approche multivariée.

MNAR Difficile à gérer sans information supplémentaire (modèles, variables,...)

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

On peut télécharger ces données sur le site de l'Insee[2].

```
insee <- read.csv("files/insee.csv", sep=";")[-15,1:4]
prop <- insee$Ensemble/sum(insee$Ensemble)*100
barplot(t(prop),yaxt="n",xaxt="n")
axis(1,line = -1,tick = F,at = (1:nrow(insee))*1.2-1/2,insee$Groupe.d.âges,las=2,cex.axis=0.4)
axis(2,at = seq(0,20,by = 3),paste(seq(0,20,by = 3),"%"),las=2,cex.axis=0.4)
```



[2] https://www.insee.fr/fr/statistiques/2381474#figure1_radio1

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

Exercice

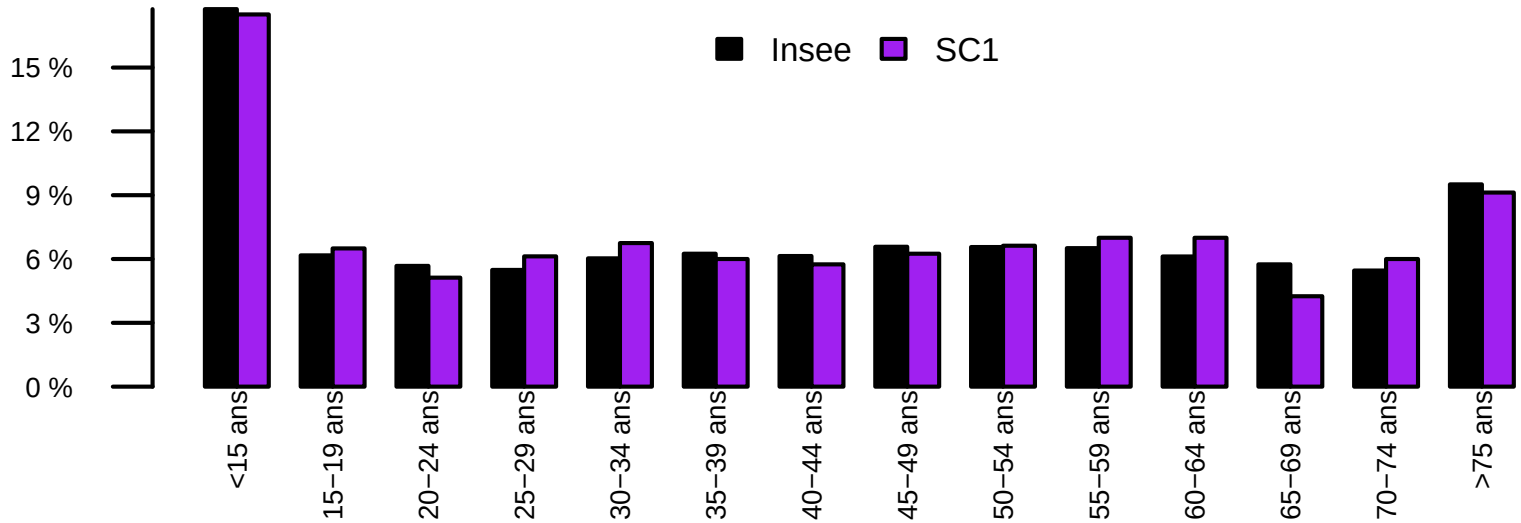
Vous devez générer un jeu de données de taille $n = 1000$ touché par un des processus de données manquantes proposés :

- $\mathbf{Sc}_1$ "10% des gens refusent de donner leur âge."
- $\mathbf{Sc}_2$ "Les moins de 5 ans ne parviennent pas à renseigner leur âge."
- $\mathbf{Sc}_3$ "C'est vendredi, les étudiants ne sont pas là!"

Vous discuterez des hypothèses prises et des résultats obtenus.

Exemple, l'âge de la population française selon l'Insee, $\mathbf{Sc}_1$

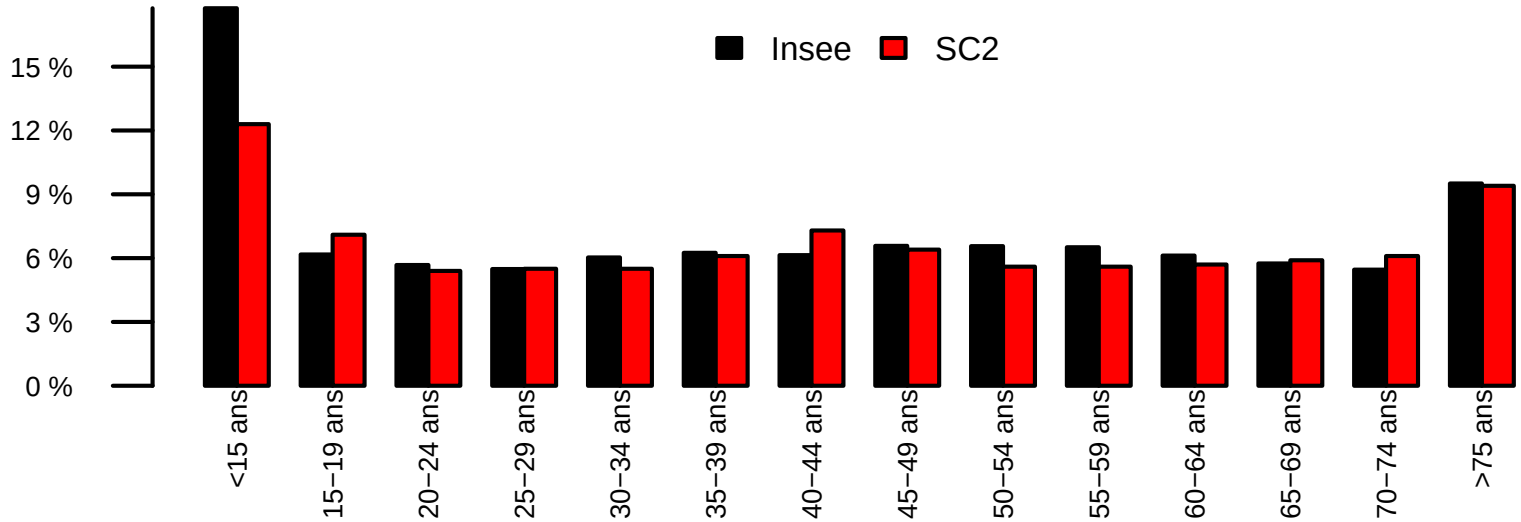
```
propNA <- 0.8
sampSC1 <- sample(x = 1:length(prop),size = propNA*n_pop,prob = prop,replace = T)
propSC1 <- table(sampSC1)/(propNA*n_pop)*100
barplot(t(cbind(prop,propSC1)),beside = T,col=c(1,"purple"),yaxt="n",xaxt="n")
axis(1,line = -1,tick = F,at = (1:nrow(insee))*3-1,insee$Groupe.d.âges,las=2,cex.axis=0.4)
axis(2,at = seq(0,20,by = 3),paste(seq(0,20,by = 3),"%"),las=2,cex.axis=0.4)
legend("top",c("Insee","SC1"),fill=c(1,"purple"),ncol=2,bty="n",cex=1/2)
```



Exemple, l'âge de la population française selon l'Insee, $\mathbf{Sc}_2$

Un tiers des individus de classe **1** est à retirer du jeu de données (hyp 1)

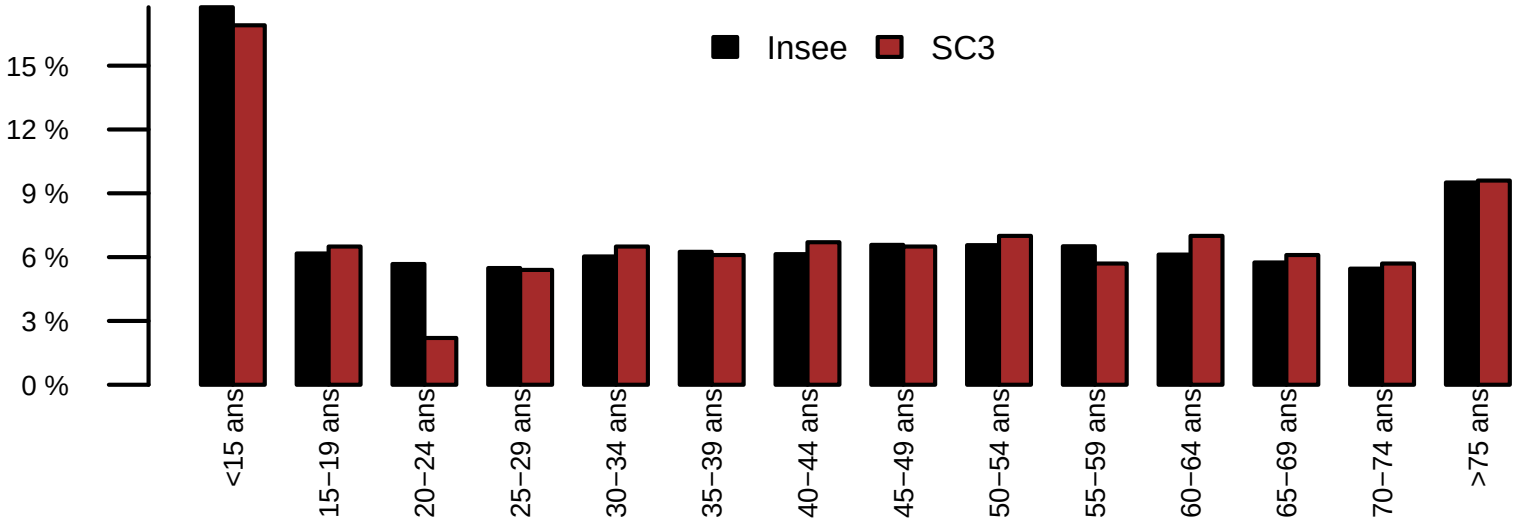
```
propSC2 <- sample(x = 1:length(prop),size = n_pop,prob = prop,replace = T)
class1 <- which(propSC2==1);class1_no5 <- class1[1:(length(class1)/3)];propSC2[class1_no5] <- NA
propSC2 <- table(propSC2)/n_pop*100
barplot(t(cbind(prop,propSC2)),beside = T,col=c(1,"red"),yaxt="n",xaxt="n")
axis(1,line = -1,tick = F,at = (1:nrow(insee))*3-1,insee$Groupe.d.âges,las=2,cex.axis=0.4)
axis(2,at = seq(0,20,by = 3),paste(seq(0,20,by = 3),"%"),las=2,cex.axis=0.4)
legend("top",c("Insee","SC2"),fill=c(1,"red"),ncol=2,bty="n",cex=1/2)
```



Exemple, l'âge de la population française selon l'Insee, $\mathbf{Sc}_3$

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

```
propSC3 <- sample(x = 1:length(prop),size = n_pop,prob = prop,replace = T)
class1 <- which(propSC3==3);class1_stu <- class1[1:(length(class1)/2)];propSC3[class1_stu] <- NA
propSC3 <- table(propSC3)/n_pop*100
barplot(t(cbind(prop,propSC3)),beside = T,col=c(1,"brown"),yaxt="n",xaxt="n")
axis(1,line = -1,tick = F,at = (1:nrow(insee))*3-1,insee$Groupe.d.âges,las=2,cex.axis=0.4)
axis(2,at = seq(0,20,by = 3),paste(seq(0,20,by = 3),"%"),las=2,cex.axis=0.4)
legend("top",c("Insee","SC3"),fill=c(1,"brown"),ncol=2,bty="n",cex=1/2)
```



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

- L'analyse est robuste à l'hypothèse MCAR $\mathbf{Sc}_1$
- $\mathbf{Sc}_2$ peut être facilement géré
- $\mathbf{Sc}_3$ est plus difficile à gérer

Section 2

De la régression à l'imputation

Modèle linéaire simple à une covariable

Soit le modèle de simulation

$$y = x + \epsilon$$

où $x \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(0, 1/4)$ et $\epsilon \perp\!\!\!\perp x$. 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 :

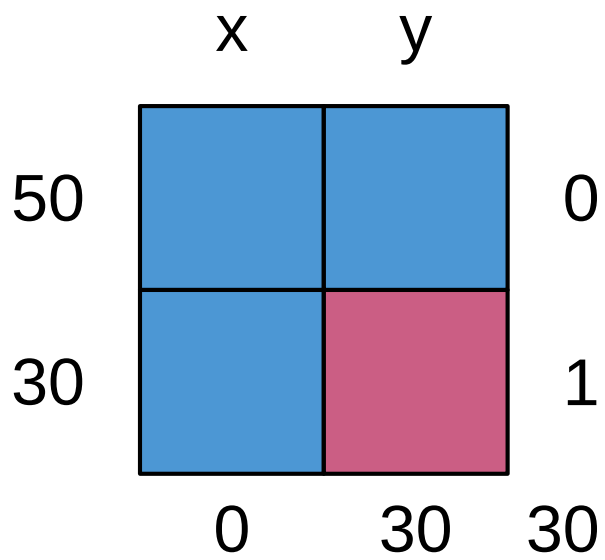
- construction d'un modèle de régression sur S, noté $\mathcal{P}$,
- estimation de la réponse de $\mathcal{P}$ à une covariable x_0, notée $\hat{y_0} = \mathcal{P}(x_0)$.

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 $(y_i)_{i=n+1,\dots,n+\tilde{n}}$.

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

```
x <- rnorm(n,0,1)  ;  y <- x + rnorm(n,0,sigma)
x_t <- rnorm(n_tilde,0,1)  ;  y_t <- rep(NA,n_tilde)
pattern_miss <- mice::md.pattern(rbind( cbind(x,y) , cbind(x_t,y_t) ))
```



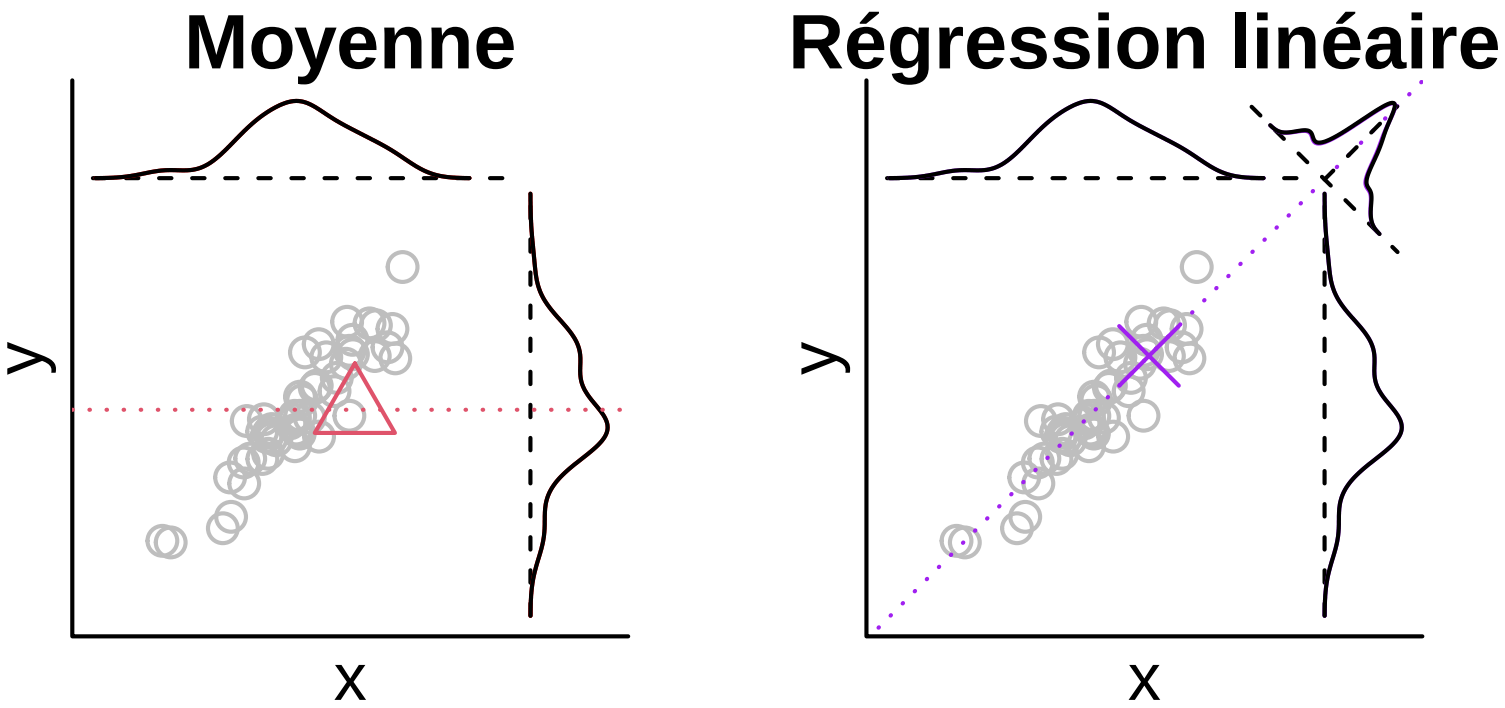
Analyse du jeu de données simulées

Comparaison de 2 méthodologies de régression :

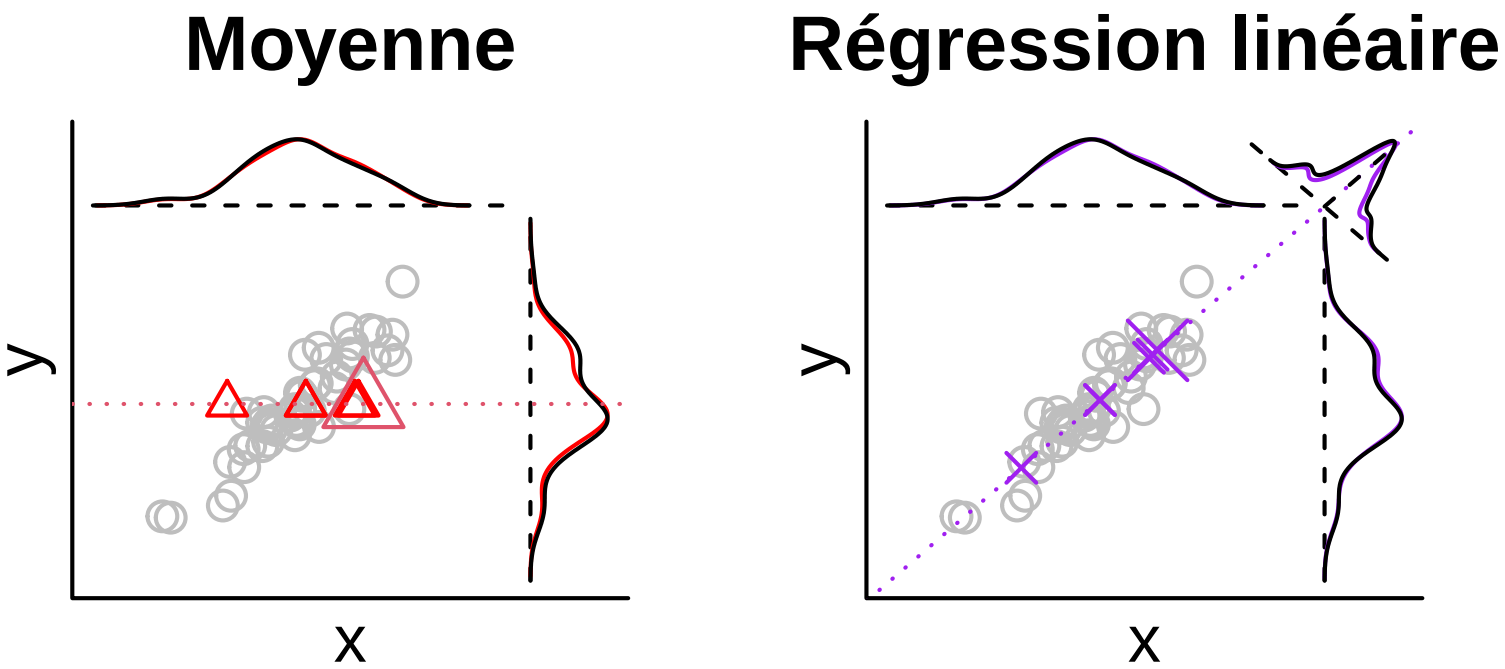
- à la moyenne,
- en utilisant le modèle de régression linéaire.

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

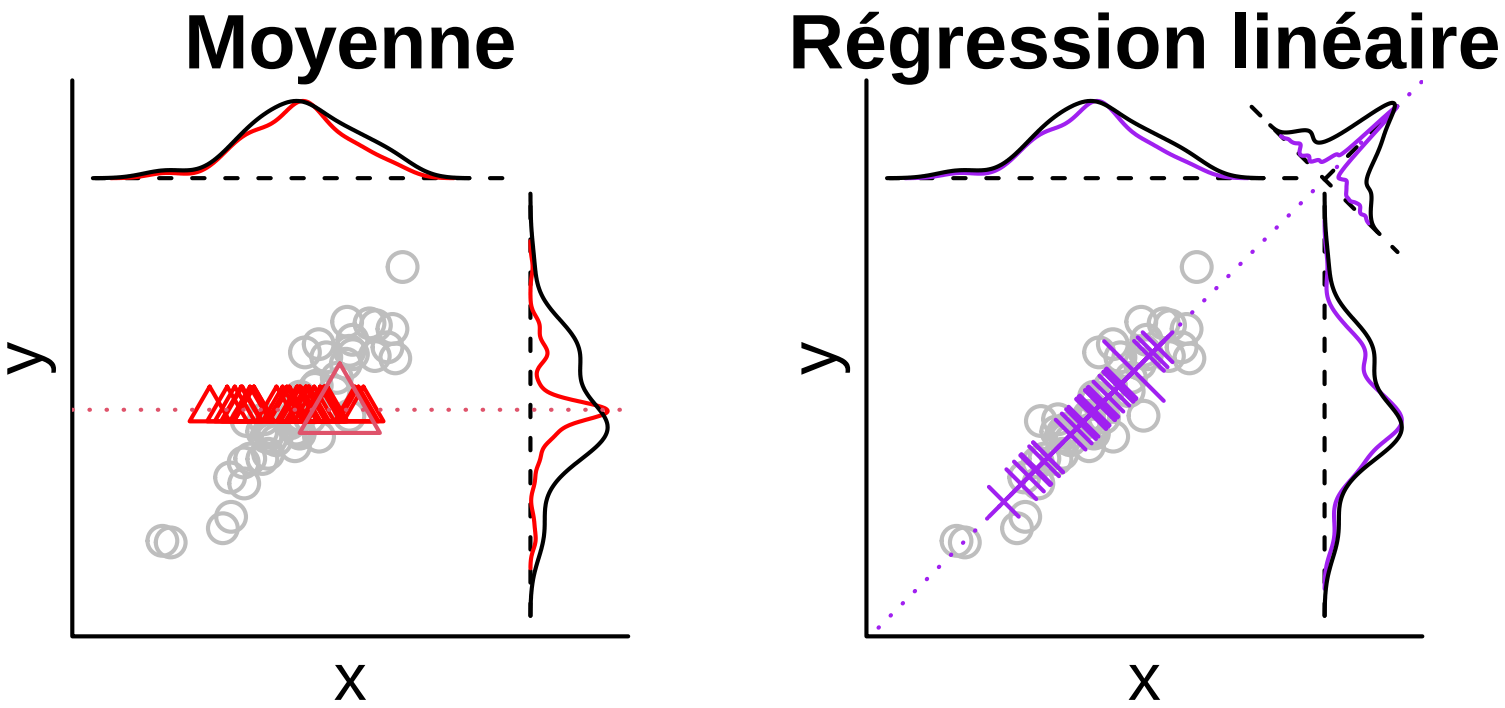
Après imputation d'une observation



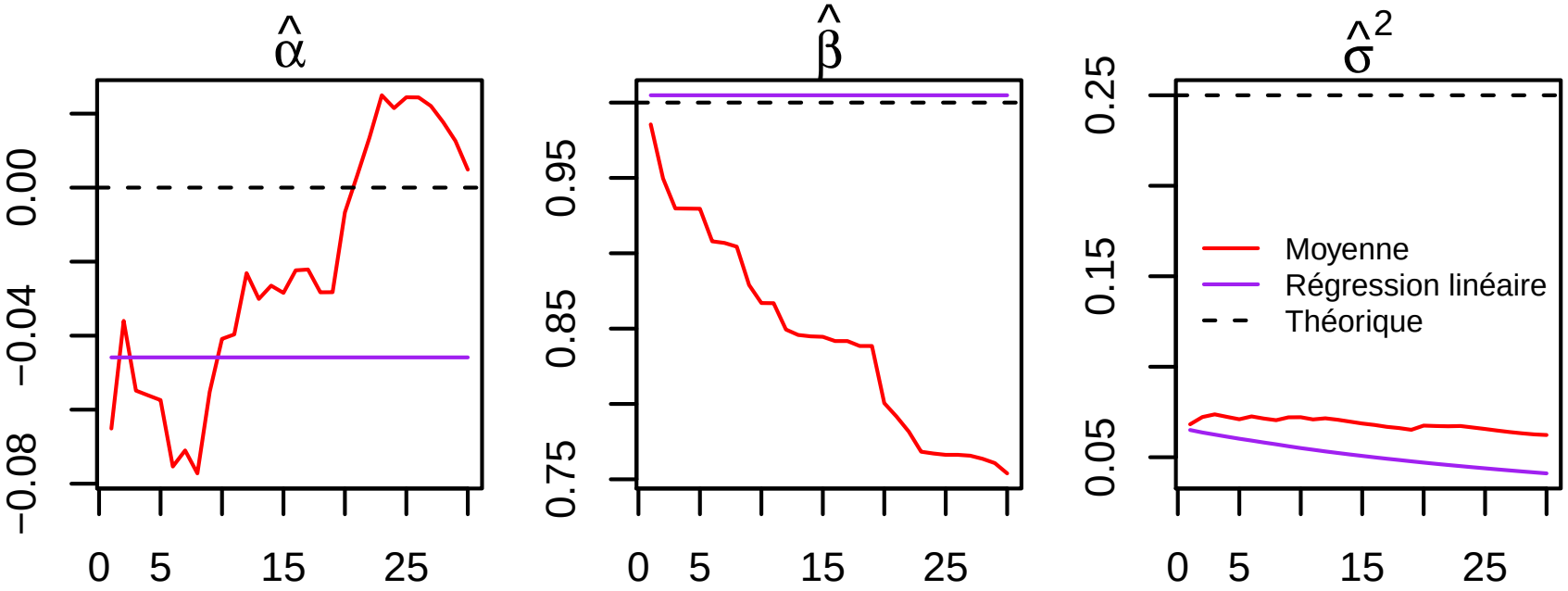
Après imputation de 5 observations



Après imputation de 30 observations



En global


α̂
β̂
σ̂²
−0.08
−0.04
0.00
0.75
0.85
0.95
0.05
0.15
0.25
0
5
15
25
Moyenne
Régression linéaire
Théorique

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 ?

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.

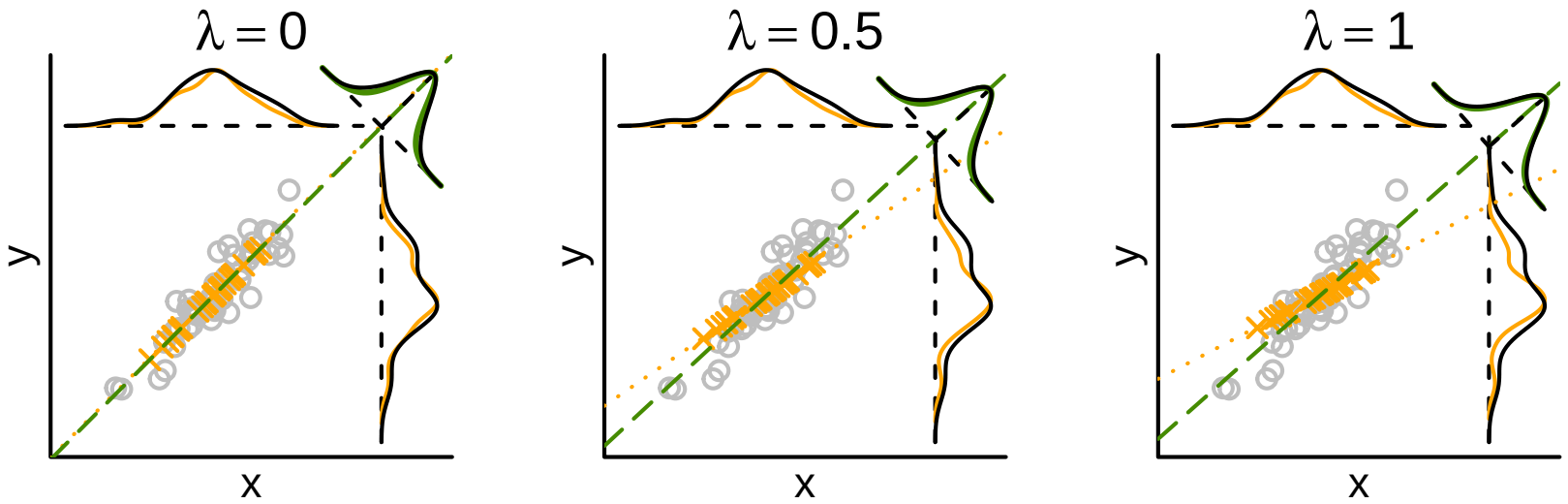
Une solution ?

Pourquoi ne pas faire de la régularisation ?

Régularisation Ridge : $\min_{\beta \in \mathbb{R}} (y - x \cdot \beta)^2 + \lambda \cdot \beta^2$

La régularisation (1)

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



λ	0	0.5	1
$\hat{\alpha}$	0.002	0.002	0.002
$\hat{\beta}$	1.005	0.926	0.885
$\hat{\sigma}^2$	0.135	0.152	0.175

$\longrightarrow$ Compromis biais-variance.

La régularisation (2)

Conclusion

La régularisation $sous - estime$ la variance de y.
Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de y.

La régularisation (2)

Conclusion

La régularisation $sous - estime$ la variance de y.
Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de y.

Conclusion Bis

Il faut utiliser une méthode qui introduise de l'incertitude sur les données imputées.

La régularisation (2)

Conclusion

La régularisation $sous - estime$ la variance de y.
Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de y.

Conclusion Bis

Il faut utiliser une méthode qui introduise de l'incertitude sur les données imputées.

Une solution ?

La régularisation (2)

Conclusion

La régularisation $sous - estime$ la variance de y.
Elle ne tient pas compte de l'incertitude sur les données manquantes pour gérer la variance de y.

Conclusion Bis

Il faut utiliser une méthode qui introduise de l'incertitude sur les données imputées.

Une solution ?

Générer M jeux de données imputés pour obtenir cette variabilité.
$\Longrightarrow$ 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

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

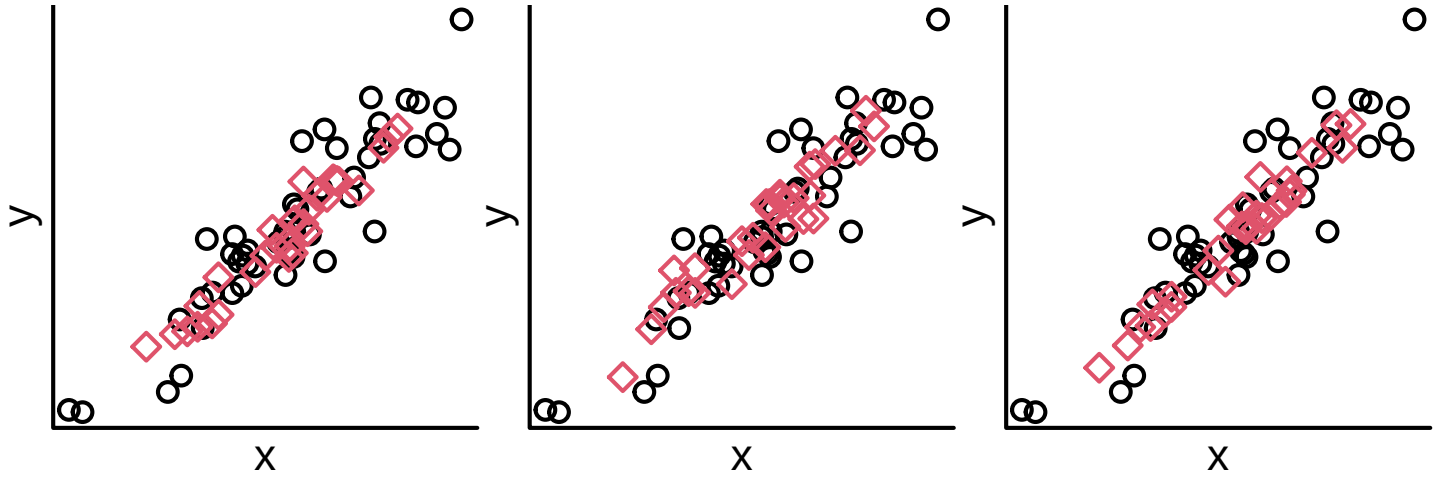
avec $\eta \sim \mathcal{N}(0, \hat{\sigma}^2)$ et $\hat{\sigma}^2$ est estimé sur S via

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

ici $p = 1$.

La régression stochastique - *improper imputation* (2)

On répette 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:

$$\hat{\alpha}_N \approx -0.051 \pm 0.044,\ \hat{\beta}_N \approx 1.002 \pm 0.046,\ \hat{\sigma}^2_N \approx 0.149$$

La régression stochastique - *improper imputation* (3)

Problème

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

Deux solutions *proper*

- Estimer M jeux de données imputés en intégrant une variabilité des paramètres à chaque fois. Plusieurs solutions ont été imaginées :
  - Tirer M paramètres sur la distribution a posteriori. Estimer M jeux de données différents en suivant ces paramètres Tanner & Wong (1987), utilisée dans **MICE** van Buuren & Groothuis-Oudshoorn (2011)
  - *Bootstraper* $(S, \tilde{S})$ pour créer M jeux de données dans lesquels il y a potentiellement des données manquantes, approche de **missMDA** Josse & Husson (2016).
- Ensuite appliquer la méthode d'analyse sur chacun des jeux de données séparément.
- Estimer les paramètres par aggrégation des M modèles.

Vers l'imputation multiple - *proper*

On obtient des estimations :

- Via bootstrap :

$$\hat{\alpha}_{\text{Boot}} \approx -0.031 \pm 0.047,\ \hat{\beta}_{\text{Boot}} \approx 1.007 \pm 0.051,\ \hat{\sigma}^2_{\text{Boot}} \approx 0.177$$

- Via la postérieure/bayésienne (une seule itération car univarié) :

$$\hat{\alpha}_{\text{Post}} \approx -0.036 \pm 0.052,\ \hat{\beta}_{\text{Post}} \approx 1 \pm 0.055,\ \hat{\sigma}^2_{\text{Post}} \approx 0.22$$

Pour rappel, dans les cas impropres :

Moyenne	$\hat{\alpha}_0 \approx 0.005 \pm 0.062$	$\hat{\beta}_0 \approx 0.754 \pm 0.066$	$\hat{\sigma}^2_0 \approx 0.31$
Régression linéaire	$\hat{\alpha}_l \approx -0.046 \pm 0.041$	$\hat{\beta}_l \approx 1.005 \pm 0.043$	$\hat{\sigma}^2_l \approx 0.135$
Régression stochastique	$\hat{\alpha}_N \approx -0.051 \pm 0.044$	$\hat{\beta}_N \approx 1.002 \pm 0.046$	$\hat{\sigma}^2_N \approx 0.149$

Règle de Rubin (*Rubin's Rule*)

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

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

où $\hat{\theta}_m$ est la valeur du paramètre estimé pour le jeu de donné d'indice m. Il vient alors :

$\hat{\sigma}^2_{IM} = \hat{\sigma}^2_W + (1+\frac{1}{M})\hat{\sigma}^2_B$,
variance imputée : $\hat{\sigma}^2_W = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^2_m$,
variance inter-imputation : $\hat{\sigma}^2_B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{IM})^2$.

Sachant que les variances par modèle sont estimées via

$$\begin{aligned}\hat{\sigma}^2_m(\hat{\alpha}) &= \hat{\sigma}^2\left(\frac{1}{n+\tilde{n}} + \frac{\bar{x}^2}{\sum_{i=1}^{n+\tilde{n}}(x_i-\bar{x})^2}\right)\\ \hat{\sigma}^2_m(\hat{\beta}) &= \hat{\sigma}^2\frac{1}{\sum_{i=1}^{n+\tilde{n}}(x_i-\bar{x})^2}\end{aligned}$$

Algorithme **EM**

EM (pour **E**stimation-**M**aximization) permet d'estimer des paramètres θ dans un modèle à vraisemblance. Soit l'espérance conditionelle, pour une itération $i > 0$:

$$Q\left(\theta, \theta^{(i)}\right) = \mathbb{E}_{x^{(mis)}}\left[\ln\left(\mathbb{P}\left(x^{(mis)}, x^{(obs)}\right)\right) | x^{(obs)}, \theta^{(i)}\right],$$

où $x^{(obs)}$ sont les données observées et $x^{(mis)}$ les données manquantes à estimer. Dans le cas normale multivarié de paramètre $\theta = (\mu, \sigma)$, cette fonction s'écrit

$$Q\left(\theta, \theta^{(i)}\right) = -\ln\left(|\Sigma|\right) - \sum_{i=1}^{n+\tilde{n}} (x_i - \mu)' \Sigma^{-1} (x_i - \mu).$$

Algorithme **EM** (2)

Résultat théorique $Q\left(\theta, \theta^{(i)}\right)$ est un fonctionnelle qui montre la vraisemblance d'un jeu de données (imputé conditionellement à $\theta^{(i)}$) pour un paramètre θ. Par construiction si

$$\theta^{(i+1)} = \text{argmax}_\theta \quad Q\left(\theta, \theta^{(i)}\right),$$

alors estimer les données manquantes conditionellement à $\theta^{(i+1)}$ est plus intéressant : $L\left(\theta^{(i+1)}|x^{(obs)}\right) \geq L\left(\theta^{(i)}|x^{(obs)}\right)$, où L est la fonction de vraisemblance. Le nouveau paramètre est donc plus vraisemblable.

Idée de l'algorithme Alterner des étapes d'estimation de $\theta^{(i)}$ et d'estimation de données manquantes.

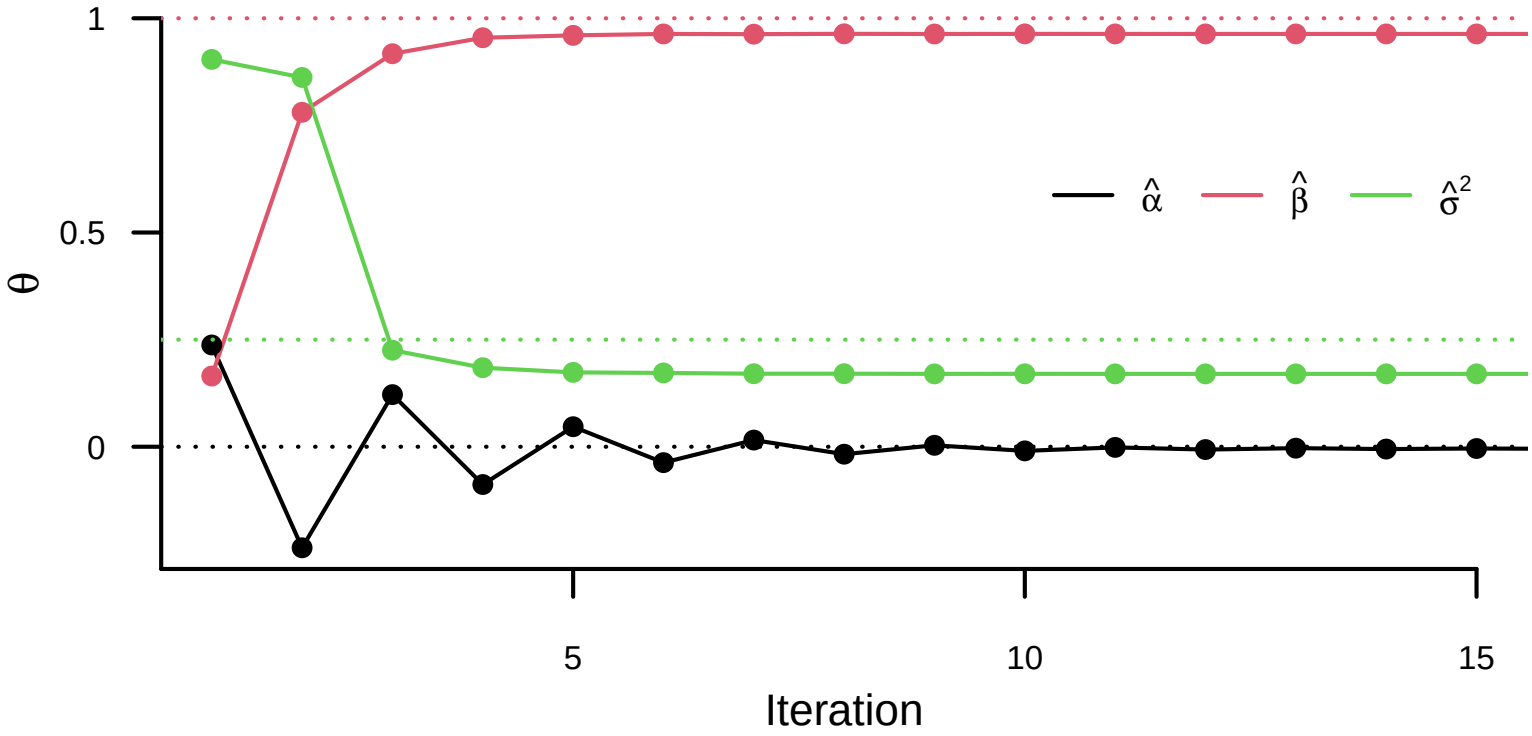
Algorithme **EM** (3)

Soit pour l'itération i en ayant initialisé $\theta^{(i)}$. On répète les deux étapes suivantes jusqu'à convergence :

Etape E Estimation des données manquantes et donc de $Q\left(\theta, \theta^{(i)}\right)$.

Etape M Maximisation de $Q\left(\theta, \theta^{(i)}\right)$ obtention de $\theta^{(i+1)}$.

Algorithme **EM** (4)



Les valeurs estimées sont alors

$$\hat{\alpha}_{\mathsf{EM}} \approx -0.005 \pm 0.044,\ \hat{\beta}_{\mathsf{EM}} \approx 0.963 \pm 0.041,\ \hat{\sigma}^2_{\mathsf{EM}} \approx 0.17$$

Section 3

Imputation et vraies données

Un passage à l'échelle nécessaire

Nous avons vu les principales charactéristiques 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 & 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 sut 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[3].

[3] 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.

Imputation en grande dimension

L'imputation en grande dimension a tendance à ne faire ressortir que les premières dimensions du jeu de donnée (nombre de composantes en ACP).

Or en grande dimension, les dimensions associées avec la réponse ne sont pas forcemment dans les premières dimensions du jeu de données.

Exemple : Ebola et données génétiques, dataset rVSV_ZEBOV Rechtien *et al.* (2017).

Il convient d'imputer en prenant en compte l'information que l'on souhaite retrouver (**y**). C'était l'objet de ma thèse et de la méthode **Koh-Lanta**, voir LORENZO *et al.* (2019).

Section 4

Conclusion

Conclusion

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

- Imbert & Vialaneix (2018), une biblio très fournie méthodo et implémentations.
- Imputation multiple & analyse factorielle, François Husson[4].
- Utilisation d'Amelia II et imputation multiple[5]

[4] http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/105543_museum_hist_nat.pdf

[5] https://cran.r-project.org/web/packages/Amelia/vignettes/intro-mi.html

Activités

- Coder l'un des algorithmes.
- Analyser un jeu de données avec (ou sans) des NA avec l'une (ou plusieurs) des méthodes discutées.

Références I

Honaker, J. & King, G. (2010) What to do about missing values in time series cross-section data. *American Journal of Political Science*, **54**, 561–581.

Imbert, A. & Vialaneix, N. (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**, 1–55.

Josse, J. & Husson, F. (2016) missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, **70**, 1–31.

Références II

Little, R.J. & Rubin, D.B. (1976) *Statistical analysis with missing data*, vol. 793, ed. John Wiley & Sons.

Lorenzo, H., Saracco, J., & Thiébaut, R. (2019) Supervised learning for multi-block incomplete data. *arXiv preprint arXiv:1901.04380*.

Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., & others (2017) Systems vaccinology identifies an early innate immune signature as a correlate of antibody responses to the ebola vaccine rVSV-zebov. *Cell reports*, **20**, 2251–2261.

Références III

Stekhoven, D.J. & Buehlmann, P. (2012) MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.

Tanner, M.A. & Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Références IV

van Buuren, S. & Groothuis-Oudshoorn, K. (2011) mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, **45**, 1–67.