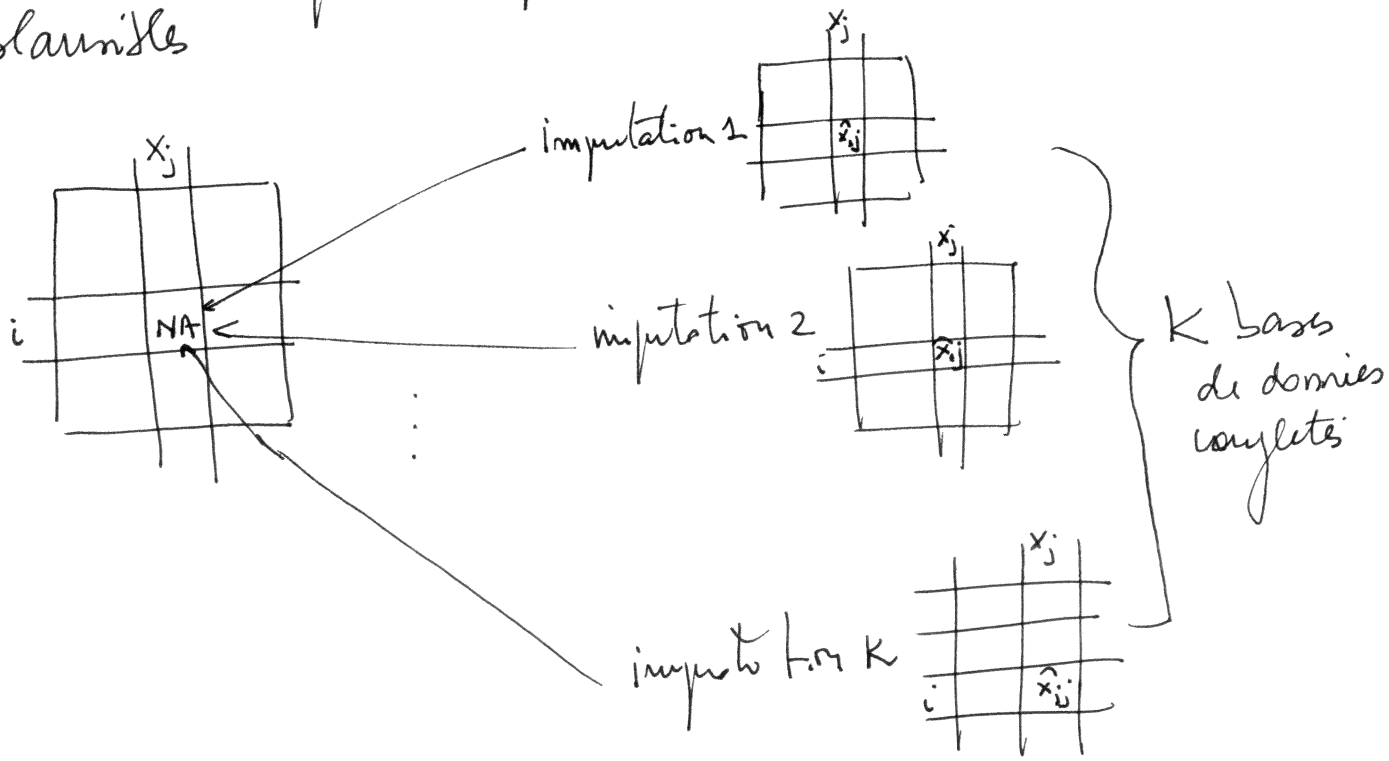
2. Imputation multiple des valeurs manquantes.

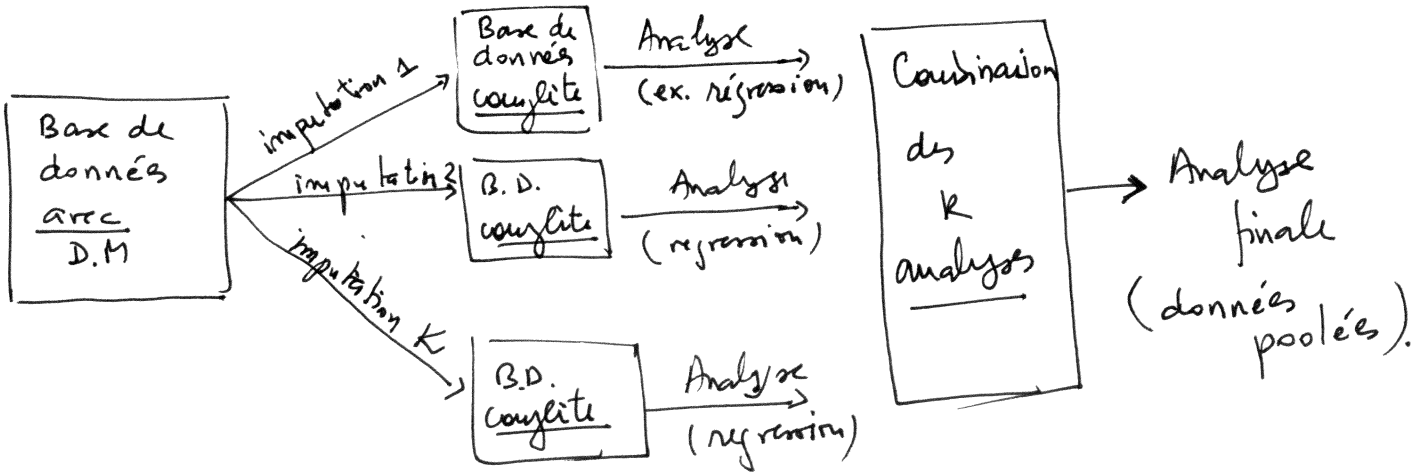
Remplacer une valeur manquante par une seule valeur estimée ⟹ pas d'incertitude sur la valeur imputée [défaut majeur]. On sous-estime aussi les variances des estimateurs en général.

Imputation multiple [Rubin 1987] : imputer la valeur manquante par deux ou plusieurs valeurs plausibles



K analyses sont réalisés sur les K bases complétés et les résultats sont ensuite mis ensemble pour obtenir une estimation globale du modèle étudié.

Processus des imputations multiples



⚠ Rubin (1987) $K = 3$ ou 5

Comment obtenir K imputations différentes pour une valeur manquante ?

• La méthode MICE (Multiple imputation Chained Equation).

Principe :

Soit X une matrice de taille $n \times p$ avec valeurs manquantes

1. On impute une première fois les données manquantes (NIPALS par exemple)
2. On prend une première variable (disons X_1) avec valeurs manquantes. On retire les valeurs imputés en 1 et on prédit ces valeurs à l'aide des autres variables $(X_2, \dots X_p)$. On obtient pour X_1 un deuxième set de imputations
3. On passe à X_2, etc. jusqu'à X_p.
4. On repete 2 et 3 $(K-1)$ fois..

Avantages de MICE :

- facile à faire
- empiriquement efficace surtout sur des grands jeux de données.

Limites : - peu de fondements théoriques ⚠

Implémenté dans le package R : mice

A lire : Journal of Statistical Software

mice : Multivariate Imputation by Chained Equations in R.

by Stef van Buuren & Karin Groothuis Oudshoorn.

https://www.jstatsoft.org/article/view/v045i03/v45i03.pdf

fonctions à regarder : md.pattern() (structure des d. m)
mice() (imputations multiples)
complete()
pool()

En SAS :

- proc means ... nmiss ... ;
- proc mi .... nimpute
  var ...
  ods select misspattern;
- proc glm
  by _imputation_ (analyses on chaque base imputée)
- Proc mianalyze ("pool" les résultats).

Exemple au II.

LA RÉGRESSION LINÉAIRE :

"Régularisation"

Y = variable aléatoire (scalaire) : la réponse quantitative

$X = (X_1, X_2, \dots, X_p)$ = variables aléatoires quantitatives : prédicteurs

Régression :

$$y = f(X_1, X_2, \dots X_p) + \varepsilon$$

Régression linéaire :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

$$\hat{y} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Estimation des coefficients de régression :

- moindre carrés ordinaires (MCO)

ou, équivalent

- maximum de vraisemblance

$(Y | X_1 \dots X_p \sim \mathcal{N}(\mu, \sigma^2))$.

Estimation à partir des données :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & \\ x_{n1} & & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

la matrice du "design"

Alors, si les variables x_j sont centrées ($\bar{x}_j = 0$)

$$\begin{cases} \hat{\beta} = (X^T X)^{-1} X^T y \text{ et} \\ \hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y \end{cases}$$

Remarque : si les variables x_j ne sont pas centrées, alors

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_p \bar{x}_p).$$

On va considérer dans la suite que les variables sont centrées (cela simplifie la présentation).

C'est quoi "la régularisation" de la régression linéaire ?

Régularisation = trouver une méthode pour résoudre les problèmes d'estimation.

Quels problèmes en régression linéaire ?

$V^{-1} = (X^T X)^{-1}$ n'existe pas !

Cette situation apparaît au moins dans deux cas :

- $n \leq p$ (peu d'observations par rapport au nb. de variables).

- multicolinéarité : variables redondantes dans le modèle

$\exists \alpha_1, \alpha_2, \dots, \alpha_p$ constantes pas toutes nulles t.q

$\alpha_1 X_1 + \dots \alpha_p X_p = 0$

Détecter ces deux problèmes est assez facile :

- multicolinéarité : à l'aide de l'ACP.

$$Vu = \lambda u$$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0 \quad , \quad \lambda_i \in \mathbb{R}_+$$

$$\downarrow u_1 \quad \downarrow u_2 \quad \downarrow u_p \qquad u_i \in \mathbb{R}^p$$

S'il existe q tq $\lambda_q = 0$ alors :

$q \in 1, \dots p$.

$c_q = X_1 u_{q,1} + \dots + X_p \cdot u_{qp}$ est telle que

$\mathbb{V}_{ar}(c_q) = \lambda_q = 0$

et donc

$$X_1 \underbrace{u_{q1}}_{\alpha_1} + \dots + X_p \underbrace{u_{qp}}_{\alpha_p} = 0$$

Observons pourquoi la multicolinéarité pose problème :

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{et}$$

$$\mathrm{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1} = \hat{\sigma}_2^2 V^{-1} \quad \text{avec}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-2} \sum (y_i - \hat{y}_i)^2$$

(variance résiduelle)

Si multicolinéarité, alors

$$\mathrm{Var}(\hat{\beta}_j) \longrightarrow \infty \qquad \forall j = 1 \dots p.$$

et comme conséquence : $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$

n'est pas significatif

$$T = \frac{\hat{\beta}_j}{\sqrt{V(\hat{\beta}_j)}} \simeq 0 \quad (\text{p-value} \simeq 1) !$$

Donc aucune variable n'est pas associée à y.

Démonstration du fait que

$$\mathrm{Var}(\beta_j) \longrightarrow \infty \quad \text{si } V^{-1} = (X^T X)^{-1}$$

n'existe pas.

(On dit aussi que V est mal-conditionnée).

Par la formule de reconstitution des données en ACP

$$X = c_1 \cdot v_1^T + c_2 \cdot v_2^T + \cdots + c_p \cdot v_p^T$$

on a que l'espace linéaire engendré par les colonnes de X (donc, par les variables X_j, $j = 1 \cdots p$) est le même que celui engendré par les composantes principales ; Donc, la régression de Y sur $\{X_1, \ldots, X_p\}$ est équivalente à la régression linéaire de Y sur $\{c_1, \ldots, c_p\}$.

$$\hat{Y} = \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

$$= \hat{\gamma}_1 c_1 + \cdots \hat{\gamma}_p c_p \quad \text{avec la}$$

relation

$$\boxed{\beta_j = \sum_{k=1}^{p} \hat{\gamma}_k \cdot \hat{u}_{kj}}$$

Or, puisque les variables $\{c_k\}_{k=1\dots p}$ sont non-corrélées, on a :

$$\mathrm{Var}(\hat{\gamma}_k) = \frac{\sigma^2}{n-p-1} \cdot \frac{1}{\lambda_k}$$

[Remarque : la matrice V de variance-covariance des $\{c_1, \dots c_p\}$ est

$$V = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

$$\Downarrow$$

$$V^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_p} \end{pmatrix}$$

et donc, comme

$$\hat{\beta}_j = \hat{\gamma}_1 \cdot u_{1j} + \hat{\gamma}_2 \cdot u_{2j} + \dots + \hat{\gamma}_p \cdot u_{pj}$$ il suit que

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{n-p-1} \cdot \left[u_{1j}^2 \mathrm{Var}(\hat{\gamma}_1) + u_{2j}^2 \cdot \mathrm{Var}(\hat{\gamma}_2) + \dots + u_{pj}^2 \mathrm{Var}(\hat{\gamma}_p) \right]$$

$$= \frac{\sigma^2}{n-p-1} \cdot \sum_{k=1}^{p} u_{k,j}^2 \cdot \frac{1}{\lambda_k}$$

Donc si $\lambda_q \simeq 0$, $1 \le q \le p$, alors

$$\frac{1}{\lambda_q} \longrightarrow \infty \text{ et donc}$$

$$\mathrm{Var}(\hat{\beta}_j) \longrightarrow \infty.$$

(fin démonstration).

Ce qu'il faut donc retenir est que

la présence des valeurs propres $\lambda_q \simeq 0$

dans l'ACP (synonime de multicolinéarité)

introduit une forte instabilité des coefficients

de la régression de Y sur $\{X_1 \dots X_p\}$.

Remarque : si $n < p$ alors évidemment,

comme $(X^T X)$ est de taille $p \times p$, dans

ce cas $\lambda_q = \lambda_{q+1} = \dots = \lambda_p = 0$, $q = n$.

car le rang $(X^T X)$ est au plus n.

→ donc, dans ce cas multicolinéarité "forcée"

(pas naturelle).

⚠ Que faire si multicolinéarité ?

Solution : dans la formule de reconstitution

des données (page 46), garder que

les composantes avec $\lambda_k >> 0$.

(composantes explicatives des données).

Que faire si multicolinéarité ?

Solutions : - faire un choix des variables : (sélection)

Vu en GIS3 :
- AIC, BIC, ...
- mais aussi, choix basé sur une bonne connaissance des variables. ("je veux dans le modèle telle et telle variable ...")

Régression sur les composantes principales :
- garder uniquement certaines composantes principales bien explicatives des données

La régression sur les composantes principales

(Principal components regression : PCR)

Idée : garder dans la formule de reconstitution des données que les composantes importantes.

Problème : les composantes "importantes" ne sont pas forcement celles les plus corrélées à Y !

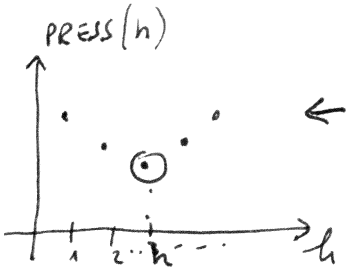
En effet, si on garde les $h < p$ comp. princ.

$$R^2(Y, \gamma_1 c_1 + \dots + \gamma_h c_h) = R^2(Y, c_1) + \dots + R^2(Y, c_h).$$

[**Rappel** : le critère de MCO revient à chercher $\beta_1 \dots \beta_p$ tq $R^2(Y, \beta_1 X_1 + \dots + \beta_p X_p)$ soit maximal

Choix des composantes principales :

- methode pas-à-pas de sélection (stepwise)



← choix de h (si les premiers h sont choisies) : par validation croisée (minimum PRESS).

package R : pls, fonction "pcr".

Le choix des composantes principales est donc un compromis entre :

- l'ajustement de Y (choisir des composantes fortement corrélées avec Y)

et

- expliquer l'information dans X (choisir des composantes fortement explicatives de X (grands λ_i)).

On peut dire aussi que ce compromis est celui entre l'ajustement (goodness of fit) et la stabilité (robustesse) des coefficients (modèle).

Pas toujours facile à faire ⚠

La regression PLS (Partial Least squares) apporte une réponse (parmi d'autres) à cela.

⚠ Voir une application de pcr sur le fichier car.txt