Méthodes de Régression Avancées

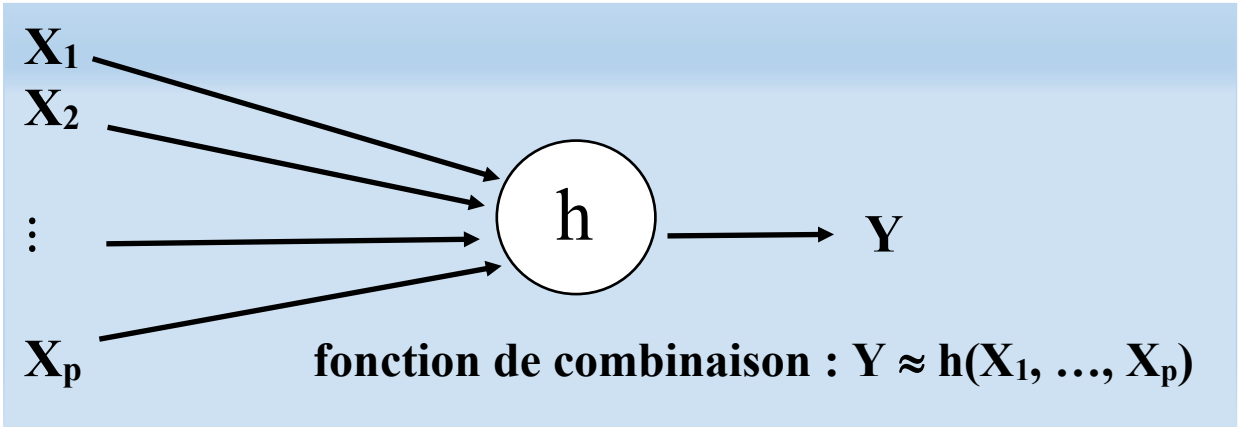
UP2 – Apprentissage statistique
Majeure Science des Données 2025-2026

Contenu

Soit **Y variable quantitative continue** que l'on cherche à **approcher** par une fonction déterministe **h** des variables $\mathbf{X_1, \ldots, X_p}$, ces dernières ayant un **lien** plus ou moins fort avec Y - **problème d'approximation ou de régression : $Y = h(X_1, \ldots, X_p) + \varepsilon$**

Le but ultime est d'utiliser $X_1, \ldots, X_p$ pour prédire la valeur de Y mais il faut auparavant construire ou « apprendre » la fonction h sur un jeu de données (**entraînement** du modèle). Cet exercice est appelé **apprentissage supervisé** vu avec les **arbres de régression (random forests, UP2)** ou à venir avec **les machines à vecteurs support (SVM) et les réseaux de neurones (UP3)**.

Langage moderne de l'**apprentissage machine** : Y variable de sortie (**output**) ou réponse et les variables $X_1, \ldots, X_p$ variables d'entrée (**inputs**) :

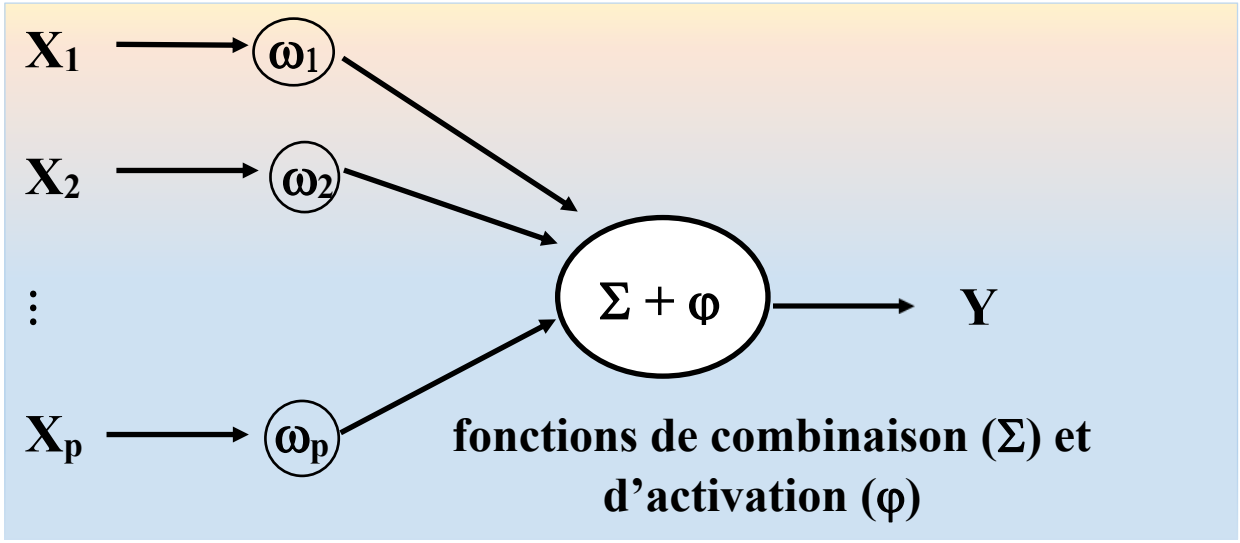


Langage de **l'apprentissage statistique** :

- Y variable à expliquer ou à prédire
- X_1, …, X_p variables explicatives ou prédicteurs

Exemple de la régression linéaire (h affine) : $\mathbf{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon}$

Vision « machine learning » de la régression linéaire :



Neurone formel

Modèle linéaire : $Y = h(X_1, \ldots, X_p) + \varepsilon \approx h(X_1, \ldots, X_p)$ où $h(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Données : $1 \le i \le n$, $y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i$ où ε_i résidu (théorique)

Estimation de β par minimisation du critère des moindres carrés (MCO ou OLS) :

$$\hat{\beta}_{MCO} = \underset{\beta}{\operatorname{argmin}} \ (J(\beta))$$

où $$\mathbf{J(\beta) = \sum_{i=1}^{n} [\, y_i - (\beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)}) \,]^2 = \| y - X\beta \|_2^2}$$

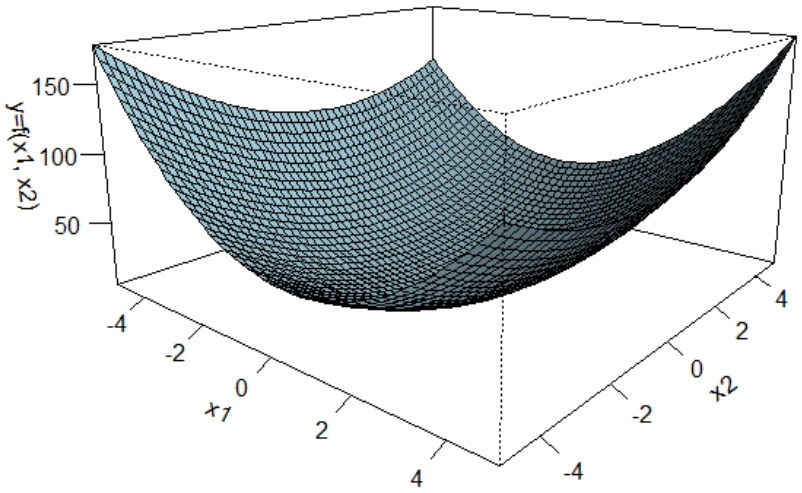
J est un critère quadratique ≥ 0 (forme quadratique positive) :

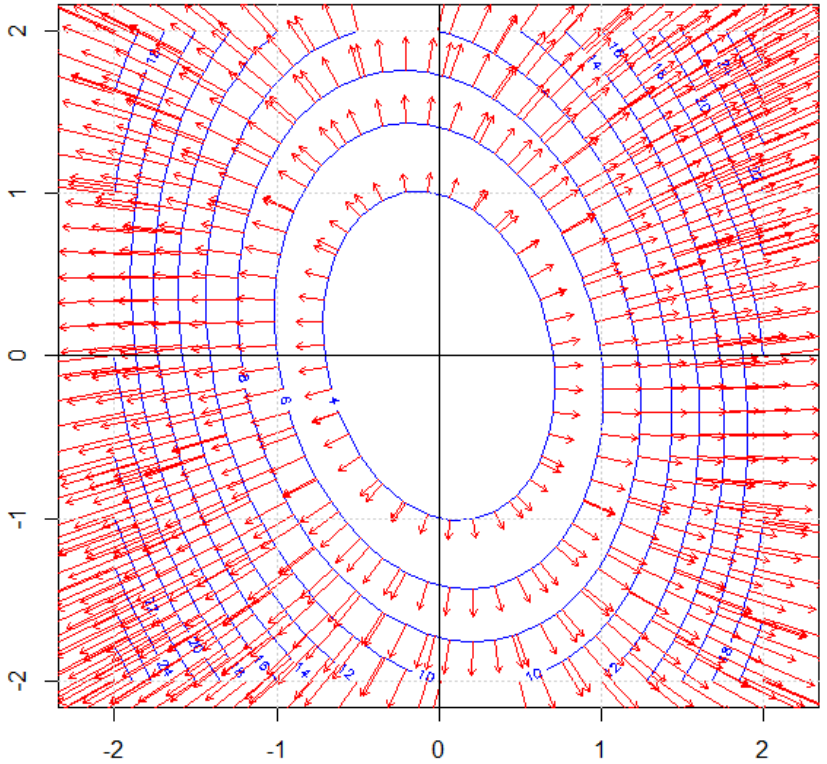
$$J(\beta) = \beta'(X'X)\beta - 2(X'y)'\beta + y'y \ge 0$$

On a $$\nabla J(\beta) = 2(X'X)\beta - 2X'y$$

☞ on peut utiliser l'algorithme du gradient classique (de la plus grande pente)

Fonction quadratique en dimension 2 et champ de vecteurs gradient





Deux limitations majeures de la Régression Linéaire Multiple usuelle (MCO ou OLS) :

- **n > p+1** et X de rang plein (p+1), alors unique minimum $\hat{\beta}$ vérifiant

$$\nabla J(\hat{\beta}) = 0 \Leftrightarrow (X'X)\beta = X'y \quad \text{(équations normales)}$$

X de rang plein signifie que les colonnes de X forment une famille libre de (p+1) vecteurs de $\mathbb{R}^n$, donc **X′X inversible** et J est une **forme quadratique définie positive**, donc une fonction **strictement convexe** (on peut toujours se ramener à X de rang plein quitte à supprimer des colonnes ou prédicteurs)

La solution peut être **instable** (**sensible à y**) ou, du point de vue statistique, la variance de $\hat{\beta}_j$ importante pour certains prédicteurs $X^{(j)}$, c'est le **phénomène de multi-colinéarité** des prédicteurs :

☞ se pose alors un **problème dit de régularisation** ou de **réduction de la variance (cf. TP1)**

- **p+1 ≥ n :** trop de paramètres, on peut ajuster de manière parfaite les données (en supposant X de rang n), ce qui est le cas extrême du **sur-apprentissage**

☞ se pose néanmoins la question de savoir s'il existe un sous-ensemble de prédicteurs (de taille faible < n) qui explique bien la réponse ! **Problème de sélection (cf. TP2)**

Retour sur le **compromis biais-variance :** T estimateur de θ, alors

$$\textbf{Risque(T)} = \textbf{Var(T)} + \textbf{Biais(T)}^2$$

Illustration classique (cf. cours 1A ou MOOC, Statistique pour l'Ingénieur, IMT)

Estimation du paramètre a d'une loi uniforme sur l'intervalle [0, a] à partir d'un échantillon $X_1, \ldots, X_n$ de taille n :

$$T_1 = 2\times\bar{X} \text{ contre } T_2 = \max(X_1, \ldots, X_n)$$

- $E(T_1) = a$ et $Var(T_1) = \frac{a^2}{3n}$
- $E(T_2) = \frac{n}{n+1} \times a$ mais

$$Var(T_2) = \frac{n\times a^2}{(n+1)^2(n+2)} \sim \frac{a^2}{n^2}$$

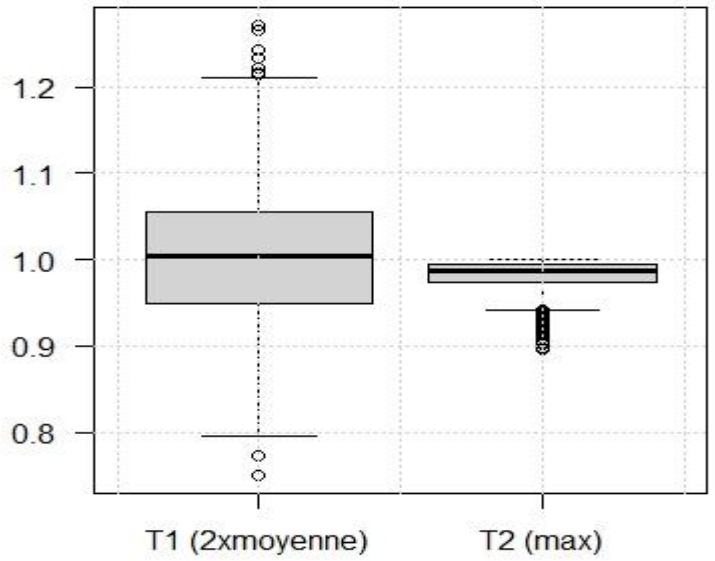


Figure 1. a = 1; n = 50 et N = 1000 échantillons indépendants simulés

Idée de la régression Ridge : contracter ou rétrécir les estimations par MCO des coefficients β (**shrinkage method**) pour diminuer la variance quitte à avoir des estimateurs biaisés $\rightarrow$ on espère diminuer le risque ☺

Critère d'ajustement Ridge (critère régularisé ou pénalisé) :

$$\mathbf{J(\beta) = \| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2} \qquad \text{avec } \boldsymbol{\lambda} \textbf{ paramètre de régularisation} > 0$$

et $\| \beta \|_2^2 = \beta_1^2 + \ldots + \beta_p^2$ donc **pas de pénalité sur l'intercept β_0 !**

☞ **on peut supposer les données centrées donc plus d'intercept β_0 dans le modèle**
☞ **il faut réduire les variables explicatives (pour que la pénalité soit insensible aux changements d'échelle)**

Minimum de J : $\hat{\beta} = \hat{\beta}(\lambda) = (X'X + \lambda I_p)^{-1} X'y$

- **$\lambda = 0$: critère usuel de la régression linéaire (MCO ou OLS)**

$$\hat{\beta}(0) = \hat{\beta}_{MCO} = (X'X)^{-1} X'y$$

- $\lambda \uparrow +\infty$: $\hat{\beta}(\lambda) \rightarrow 0$

Effet de la régularisation (régression de crête)

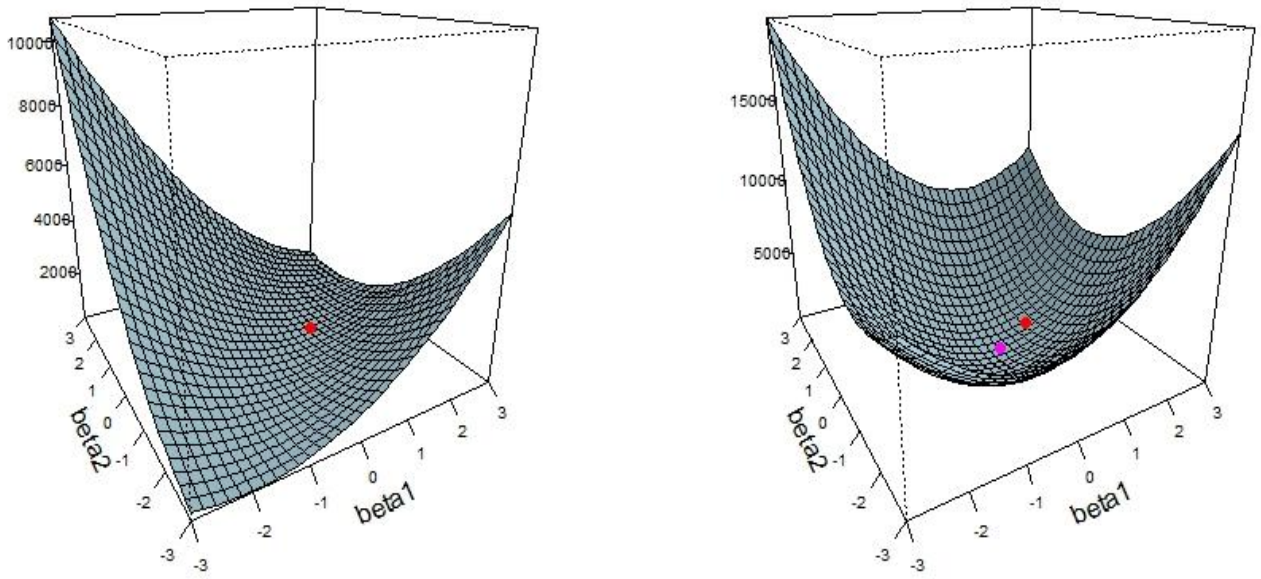


Figure 2 (p = 2). Critère $J(\beta_1, \beta_2)$ usuel à gauche (non régularisé) et critère régularisé à droite ($\lambda = 500$)
Données simulées : $y = 0.5x^{(1)} - 0.5x^{(2)} + \varepsilon$ (cf. TP1)

Minimisation de J par

- méthode de descente : $\nabla J(\beta) = 2(X'X + \lambda I_p)\beta - 2X'y$
- résolution du système linéaire : $(X'X + \lambda I_p)\beta = X'y$

Interprétation géométrique ☺ : équivalence avec le problème suivant

$$\text{Minimiser } \mathbf{J(\beta) = || y - X\beta ||_2^2} \text{ sous la contrainte } \mathbf{|| \beta ||_2^2 \leq t}$$

- $\mathbf{t \geq || \hat{\beta}_{MCO} ||_2^2}$ $\mathbf{: \hat{\beta}(t) = \hat{\beta}_{MCO} = (X'X)^{-1} X'y}$
- $\mathbf{t \downarrow 0 : \hat{\beta}(t) \rightarrow 0}$

Correspondance entre λ et t (shrinkage coefficient)

Choix du paramètre de régularisation λ

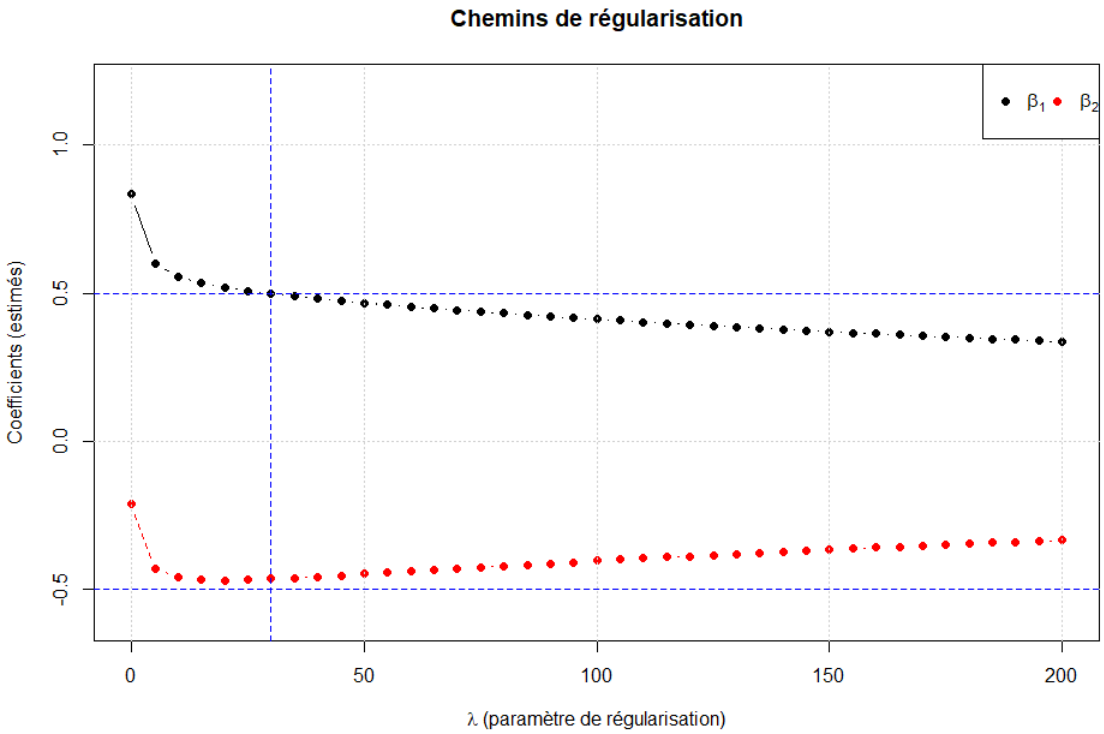


Figure 3. Profils des coefficients Ridge lorsque λ varie – Comparer avec les coefficients réels : $\beta_1 = 0.5$ et $\beta_2 = -0.5$

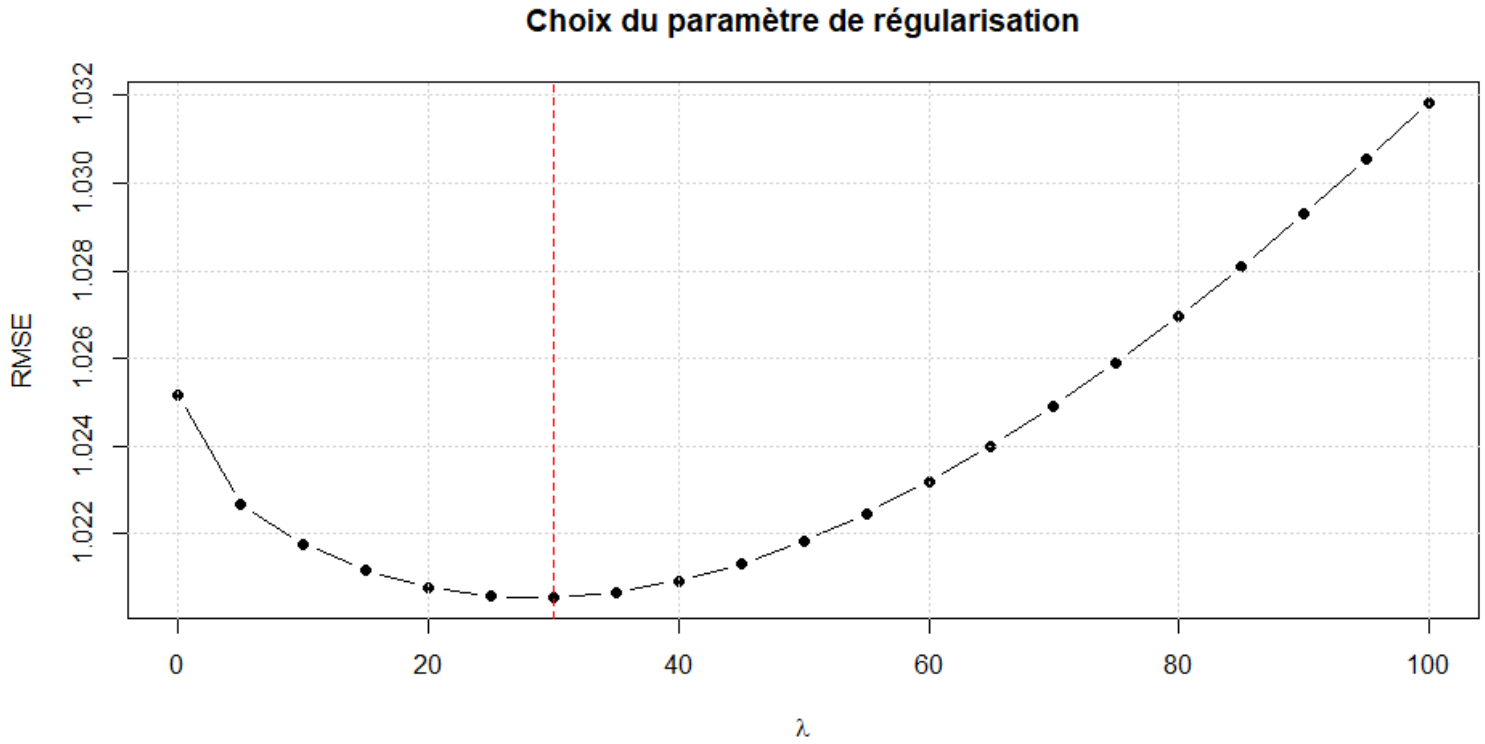
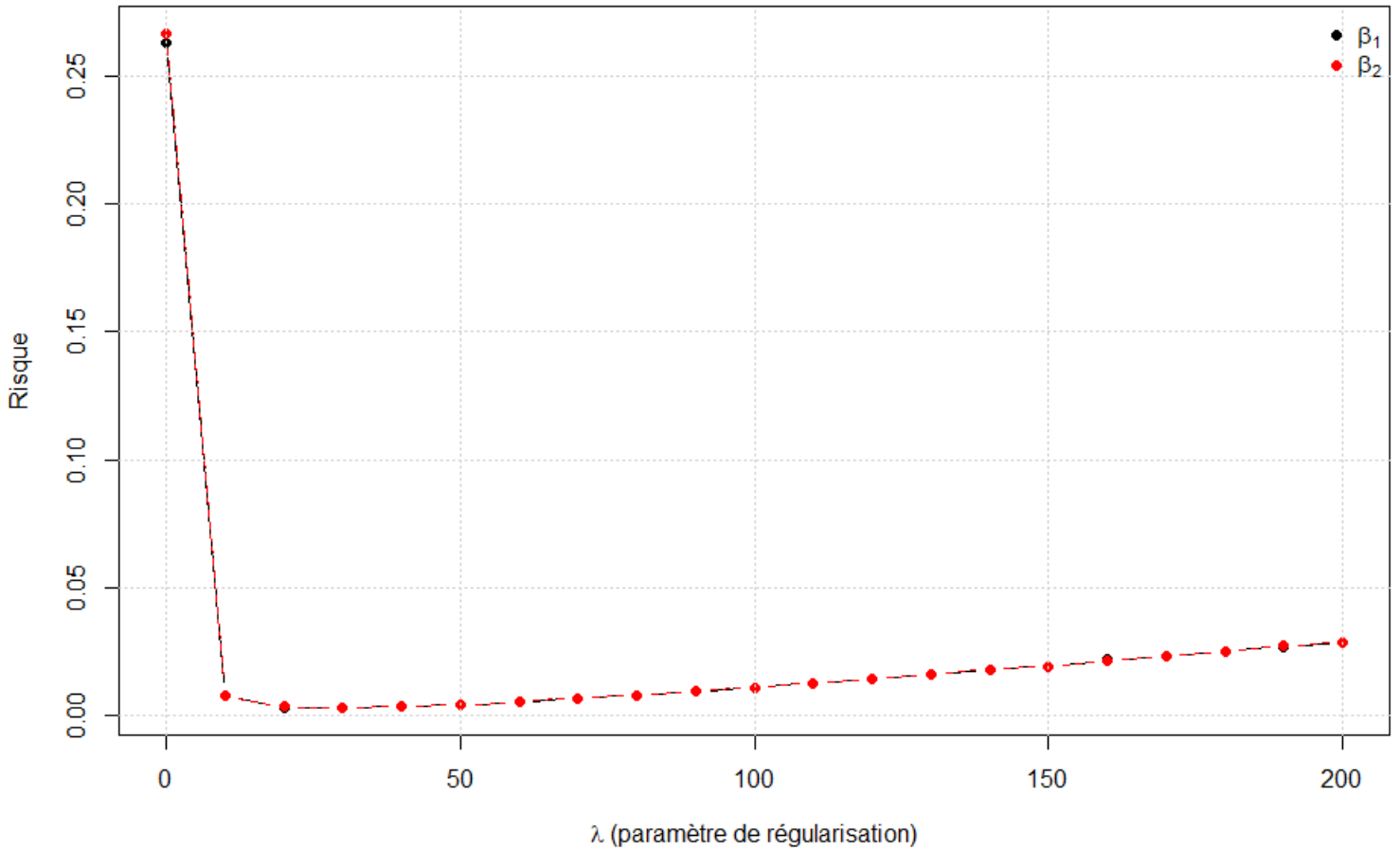


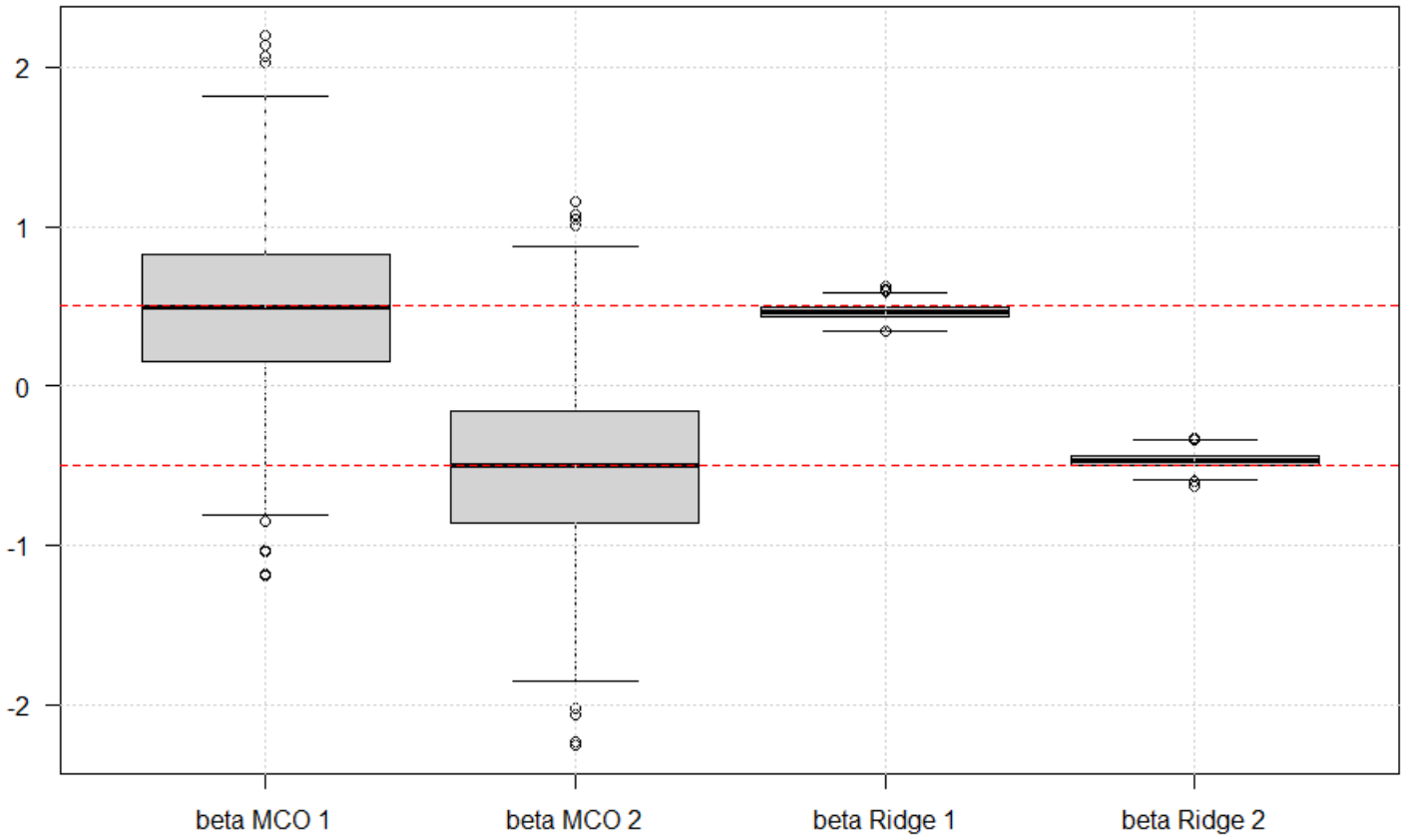
Figure 4. Choix du paramètre de régularisation
par **validation croisée** (CV) qui utilise le critère d'erreur usuel (RMSE)

$$\mathbf{RMSE} = \sqrt{\mathbf{MSE}} \text{ avec } \mathbf{MSE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} [\, y_i - \hat{y}_i(\lambda) \,]^2 = \mathbf{MSE}(\lambda)$$

Compromis biais-variance



Distribution estimateurs MCO et Ridge



Critère d'ajustement LASSO :

$$\mathbf{J(\beta) = \| y - X\beta \|_2^2 + \lambda \| \beta \|_1} \qquad \text{avec } \boldsymbol{\lambda} \textbf{ paramètre de régularisation} > 0$$

et $\| \beta \|_1 = | \beta_1 | + \ldots + | \beta_p |$ (Least Absolute Shrinkage ! and **Selection Operator ?**)

☞ **Tibshirani, Robert** 1996, Journal of the Royal Statistical Society - voir aussi :
The Elements of Statistical Learning, Trevor Hastie, R. T. and Jerome Friedman, 2009, Springer

Minimum de J : $\hat{\beta} = \hat{\beta}(\lambda)$

- $\lambda = 0$ **: critère usuel de la régression linéaire (MCO ou OLS)**

$$\hat{\beta}(0) = \hat{\beta}_{MCO} = (X'X)^{-1} X'y$$

- $\lambda \uparrow +\infty$ **:** $\hat{\beta}(\lambda) \to \mathbf{0}$

Interprétation géométrique ☺ : équivalence avec le problème suivant

$$\text{Minimiser } \mathbf{J(\beta) = \| y - X\beta \|_2^2} \text{ sous la contrainte } \| \beta \|_1 \leq t$$

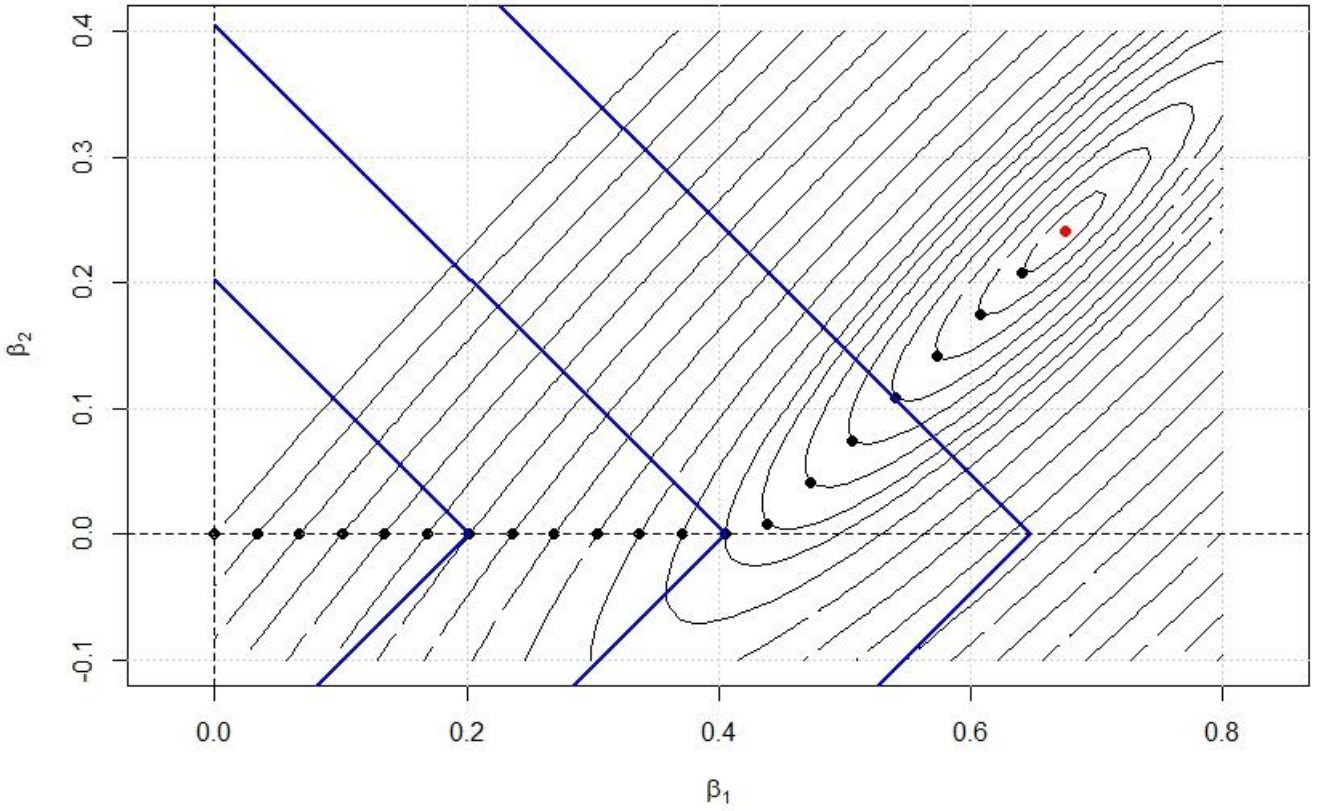


Figure 5 (p = 2). Chemin de régularisation et sélection avec le LASSO

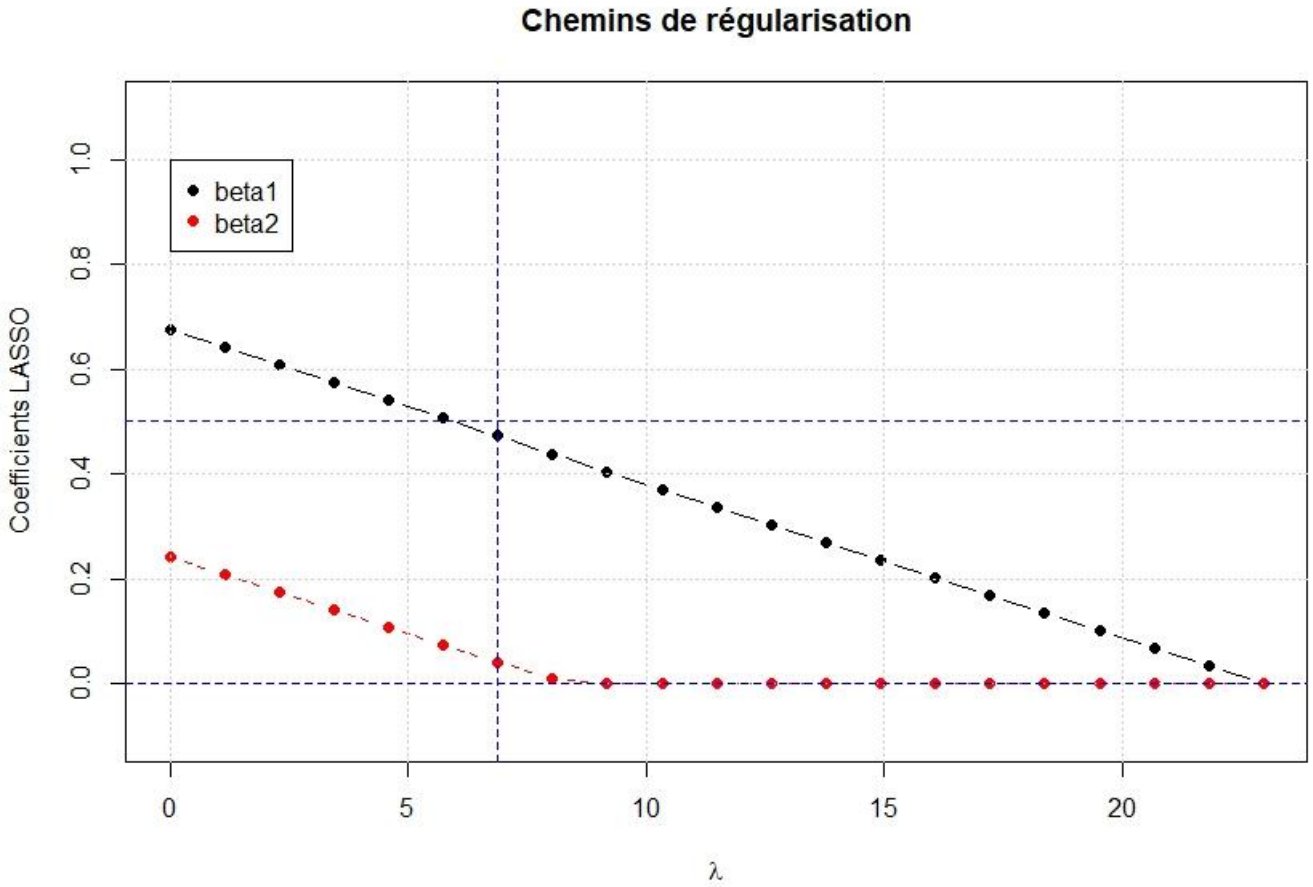


Figure 6 (p = 2). Profils des coefficients LASSO lorsque λ varie. La ligne verticale en pointillés indique la valeur choisie par **validation croisée** ($\lambda_{CV} = 6.89$)

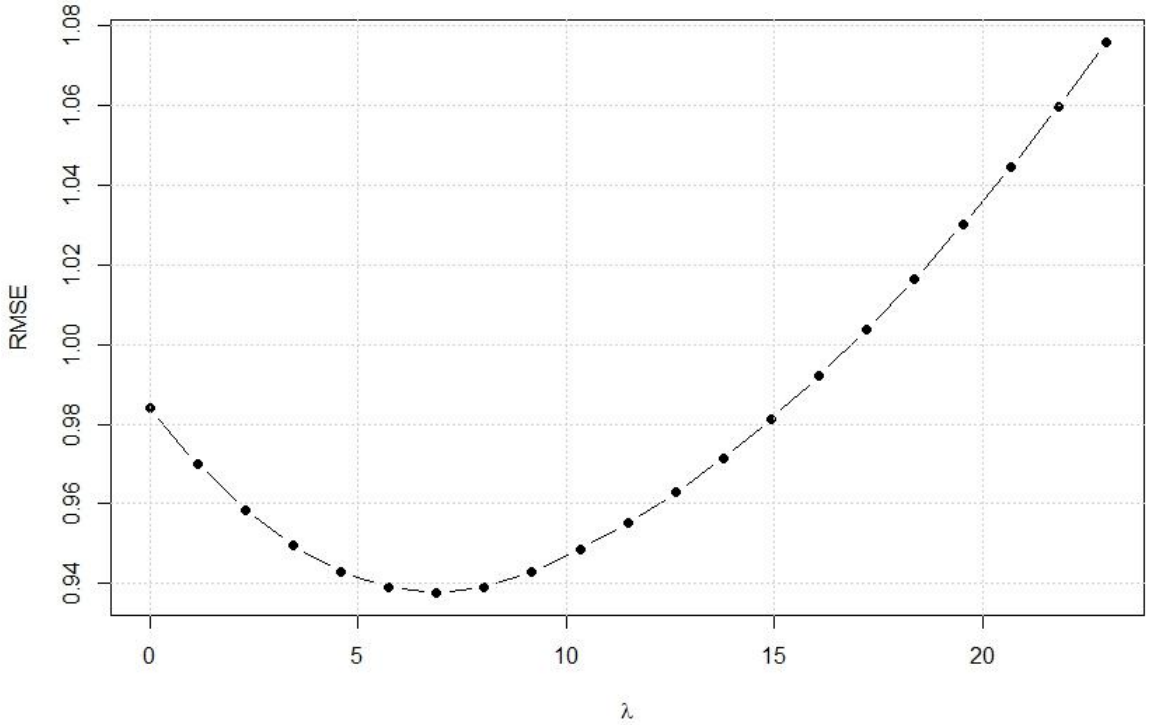


Figure 7. Choix du paramètre de régularisation
par **validation croisée** (CV) qui utilise le critère d'erreur usuel (RMSE)

Soient $Z^{(1)}$, …, $Z^{(p)}$ les **composantes principales** associées aux variables $X^{(1)}$, …, $X^{(p)}$ supposées centrées-réduites en général (ACP normée) :

- $Z^{(1)} = v_{1,1} X^{(1)} + \ldots + v_{1,p} X^{(p)}$ de variance (inertie) maximale λ_1 avec v_1 vecteur unitaire de $\mathbb{R}^p$ (espace des individus)

- $Z^{(m)} = v_{m,1} X^{(1)} + \ldots + v_{m,p} X^{(p)}$ de variance (inertie) maximale λ_{m-1} avec v_m vecteur unitaire de $\mathbb{R}^p$ orthogonal à v_1, …, v_{m-1}

Sous forme matricielle, on a encore (DVS ou SVD de la matrice X) :

$$\mathbf{X = U\Sigma V' \Leftrightarrow X = \sqrt{\lambda_1} < v_1 \mid . > u_1 + \ldots + \sqrt{\lambda_p} < v_p \mid . > u_p)}$$

On sait que $Z^{(1)}$, …, $Z^{(p)}$ correspondent à de nouvelles variables qui sont **non corrélées entre elles** $\Leftrightarrow$ **famille orthogonale de vecteurs de** $\mathbb{R}^n$ (espace des variables).

Régression sur Composantes Principales (Principal Component Regression) :

$$\mathbf{y = \theta_1 Z^{(1)} + \ldots + \theta_M Z^{(M)} + \varepsilon} \qquad \text{avec } 1 \leq M \leq p \quad (M = \text{dimension})$$

On a $$\hat{\theta}_m = \frac{< \mathbf{y} \mid \mathbf{Z^{(m)}} >}{\| \mathbf{Z^{(m)}} \|^2} \qquad \text{pour m = 1 à M}$$

d'où le prédicteur PCR :

$$\hat{y} = \hat{\theta}_1 Z^{(1)} + \ldots + \hat{\theta}_M Z^{(M)} = \hat{\beta}_1 X^{(1)} + \ldots + \hat{\beta}_p X^{(p)}$$

- M = p : on garde toutes les composantes et on retrouve le prédicteur des MCO
- Si bonnes prédictions avec M < p, on a réduit la dimension réelle de l'espace des prédicteurs
- Cette méthode est très proche de la régression Ridge qui consiste à contracter les coefficients β du modèle de régression linéaire usuel (on met à 0 les coefficients des composantes principales d'ordre M+1, …, p)
- Détermination du « paramètre » M par validation croisée
- Gros défaut : les premières composantes principales ne sont pas forcément les variables les plus corrélées avec la réponse Y, d'où la régression PLS…

L'idée de la régression par moindres carrés partiels (**Partial Least Squares Regression**) est encore de chercher à régresser sur un nombre M plus petit de composantes non corrélées de variances les plus grandes possibles mais qui soient également les plus corrélées possibles avec la réponse y ($M \leq p$, d'où les carrés partiels).

On considère encore les données centrées-réduites.

On cherche la première composante PLS sous la forme suivante :

$$W^{(1)} = w_1 X^{(1)} + \ldots + w_p X^{(p)} \text{ avec w vecteur unitaire de } \mathbb{R}^p \text{ (espace des individus)}$$

de sorte à maximiser le critère

$$\mathbf{cov(W^{(1)}, y)} \text{ ou } \mathbf{cov(W^{(1)}, y)^2 = cor(W^{(1)}, y)^2 \times var(W^{(1)}) \times var(y)}$$

On trouve (à une constante près) que

$$W^{(1)} = < y \mid X^{(1)} > X^{(1)} + \ldots + < y \mid X^{(p)} > X^{(p)}$$

On considère ensuite les variables

$$\mathbf{X^{(j)} - E_L(X^{(j)} \mid W^{(1)})} \text{ pour } 1 \leq j \leq p$$

obtenues par « orthogonalisation » des prédicteurs $X^{(j)}$ relativement à $W^{(1)}$.

Pour j fixé, $X^{(j)} - E_L(X^{(j)} \mid W^{(1)})$ est la composante du prédicteur $X^{(j)}$ non expliquée par $W^{(1)}$.

On cherche enfin dans le sous-espace de dimension (p – 1) engendré par ces variables la seconde composante PLS notée $W^{(2)}$.

On itère ce procédé, ce qui fournit une famille orthogonale : $W^{(1)}, \ldots, W^{(M)}$

Régression PLS

$$\mathbf{y = \theta_1 W^{(1)} + \ldots + \theta_M W^{(M)} + \varepsilon} \qquad \text{avec } 1 \leq M \leq p$$

On a encore $$\hat{\theta}_m = \frac{< y \mid W^{(m)} >}{\| W^{(m)} \|^2} \qquad \text{pour m = 1 à M}$$

On considère ici le problème direct de sélectionner un bon sous-ensemble de prédicteurs parmi p qui soit de taille minimale tout en étant le meilleur possible pour expliquer ou prédire la réponse y.

Il faut pour cela se donner un critère de performance (exemple simple avec le **R^2 ajusté** qui pénalise les modèles trop paramétrés) :

- **Sélection exhaustive**
- **Sélection pas à pas descendante (backward regression)**
- **Sélection pas à pas ascendante (forward regression)**
- **Sélection bidirectionnelle…**
- **…**