19/09

→ Jeu de données : $(x_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$

x_{ij} = valeur de la variable j pour l'observation i

Que veut-on faire de ces données

descriptive

→ descriptif : faire des groupes d'individus homogènes (clusters)

on parle de clustering (classification automatique ...)

méthodes : k-means, CAH ...

prédictive

→ y_i : catégorielle (ex pathologie) à expliquer

classification (supervisée)

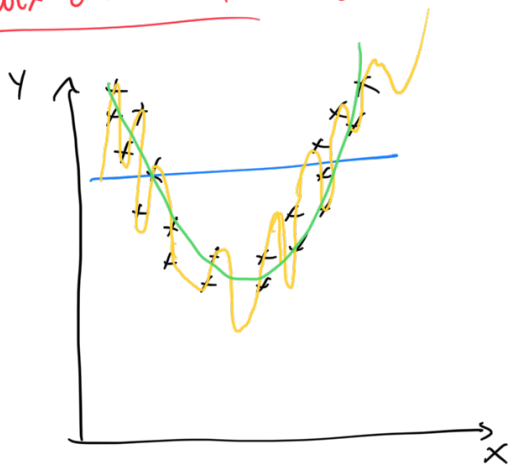
4(but prédictif, but explicatif)

→ y_i : quantitative

régression

Choix de modèles

on veut comparer des modèles de régression



$y_i = a + b x_i + \varepsilon_i$

- résidus ε_i très grands, un grand biais
- peu sensible aux fluctuations d'échantillonnage : une variance faible

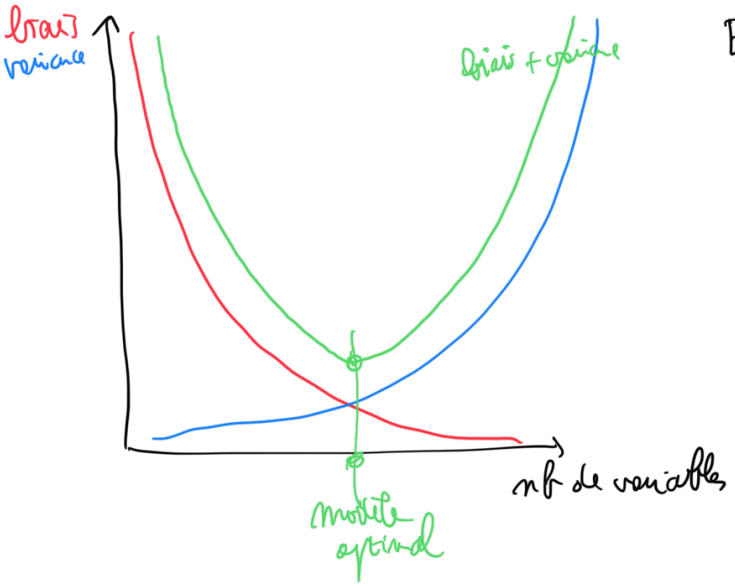
$y_i = a + b_{20} x_i^{20} + b_{19} x_i^{19} + \dots + b_1 x_i + \varepsilon_i$

- résidus très faibles : biais faible
- variance très grande

SUR-AJUSTEMENT

$y_i = a + b_2 x_i^2 + b_1 x_i + \varepsilon_i$

biais faible, variance faible] bon compromis biais variance



En pratique on va comparer les modèles

↳ critère AIC, BIC

↳ évaluer la prédiction sur des données indépendantes (qui n'ont pas servi à estimer les modèles)

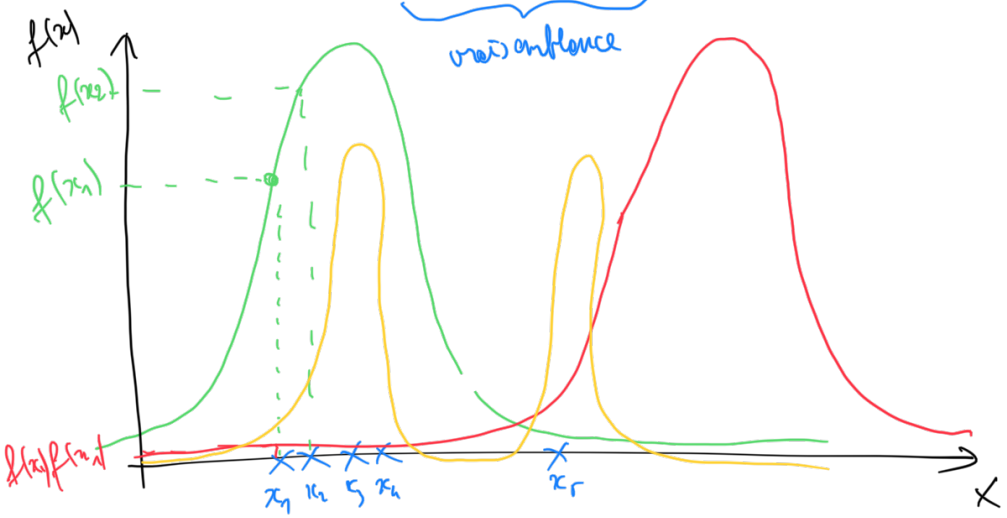
- apprentissage / test
- validation croisée.

↳ R^2 ajusté

(R^2 croît en fonction du nb de variables)

$f(x_1) \times f(x_2) \times \dots \times f(x_5) > f(x_1) \times f(x_2) \times \dots \times f(x_5)$

vraisemblance



⚠ A modèle fixé, on estime généralement les paramètres par maximum de vraisemblance

Mais on ne l'utilise pas pour comparer des modèles, car la vraisemblance croît en fonction du nombre de paramètres.

Régression généralisée

modèle $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Max de vraisemblance ≈ moindres carrés (ordinary least square OLS)

on cherche $\beta = (\beta_0, \dots, \beta_p)$ qui minimise

$$\sum_i \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2$$

Si p grand, et/ou si variable sont corrélées $\mathbb{V}(\hat{\beta}_j)$ grande
on observe des $\hat{\beta}_j$ qui peuvent prendre de grandes valeurs
suivant les fluctuations d'échantillonnage

$$\hat{\beta}^{OLS} = (X^t X)^{-1} X^t Y$$

On va donc résoudre un problème annexe en
empêchant les β_j de prendre des valeurs trop grandes

$\Rightarrow$ on cherche $\beta = (\beta_0, \dots, \beta_p)$ qui minimise
on cherche des β qui

$$\underbrace{\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_j \beta_j x_{ij} \right) \right)^2}_{\text{donne de petits résidus}} + \lambda \underbrace{\sum_j \beta_j^2}_{\text{sans prendre de valeurs trop grandes}}$$

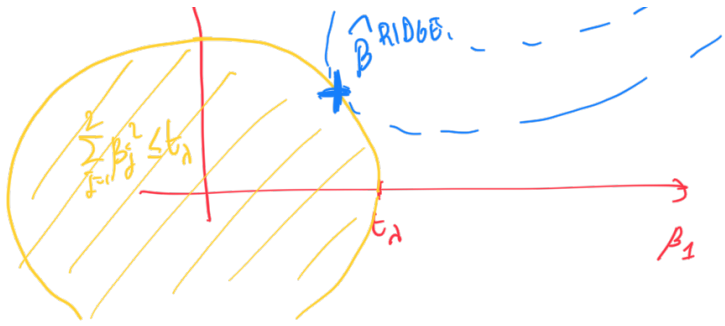
$\Leftrightarrow$ ou de façon équivalente, on cherche β

qui minimise $\sum_i \left(y_i - \left(\beta_0 + \sum \beta_j x_{ij} \right) \right)^2$ sous les contraintes $\sum_j \beta_j^2 \leq t_\lambda$

où λ (t_λ) sont des hyper-paramètres à régler.

$p=2$

RIDGE Regression



La solution est donnée par $\begin{pmatrix} \hat{\beta}_1 \\ 0 \end{pmatrix}$

$$\hat{\beta}^{RIDGE} = (X^t X + \lambda I)^{-1} X^t Y$$

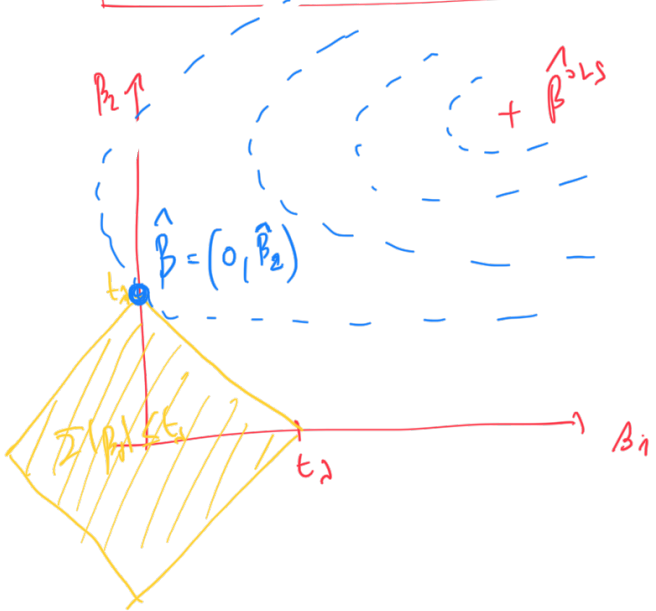
Régression régularisée / pénalisée.

Changeons la pénalité $\sum \beta_j^2$ par $\sum |\beta_j|$

$\Rightarrow$ On cherche β qui minimise

$$\sum_i \left(y_i - \left(\beta_0 + \sum_j \beta_j x_{ij}\right)\right)^2 \quad \text{sous la contrainte } \sum_j |\beta_j| \leq$$

Régression LASSO

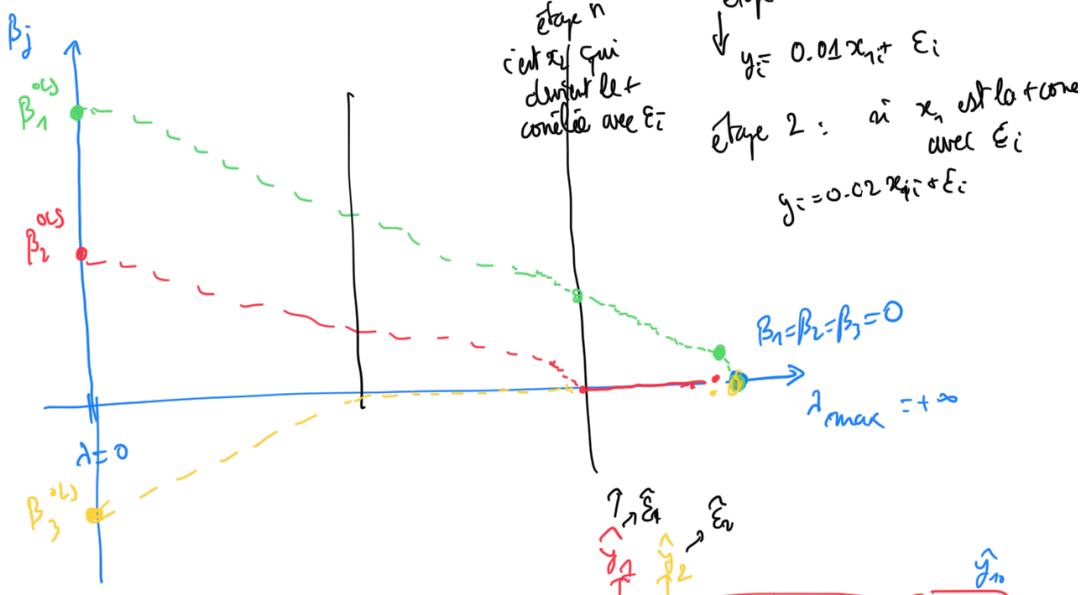


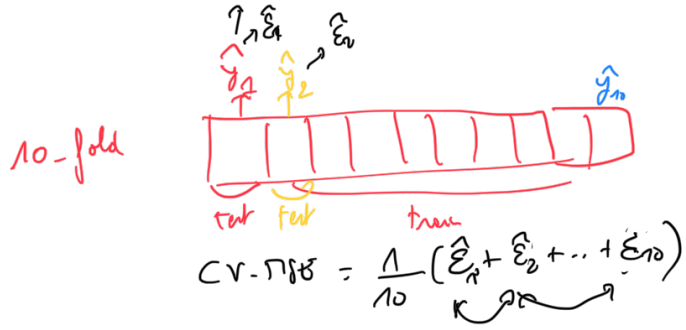
Pb → pas de solution analytique $\hat{\beta}^{LASSO}$ = ??

→ on utilisera un algorithme d'optimisation numérique qui nous donnera une

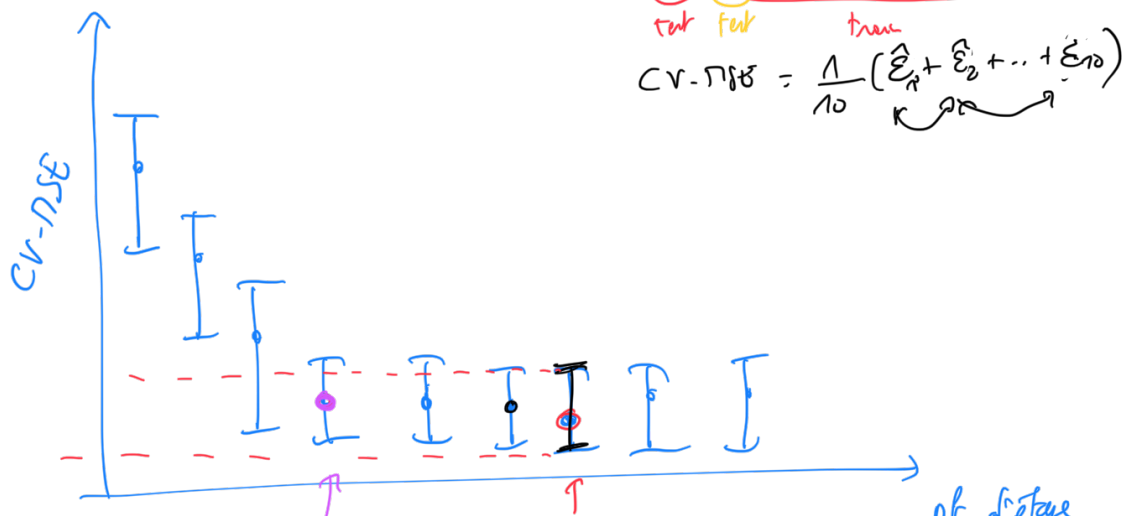
numérique que [illegible]
solution exacte de β LASSO.

Algorithme LARS





$$CV\text{-}MSE = \frac{1}{10}\left(\hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \dots + \hat{\varepsilon}_{10}^2\right)$$



20/09

- Lorsque j'ai beaucoup d'individus (lignes) dans ma base

de données, c'est un avantage d'un point de vue statistique

oui ☐ non ☐

- En présence de variables continues et catégorielles
  - ☐ je décide à utiliser des méthodes qui permettent nativement de prendre en compte ces deux types de données
  - ☐ je discrétise les variables continues en variables catégorielles
  - ☐ je transforme les données catégorielles en données quantitatives
- La présence de corrélation dans les covariables induit:
  - ☐ un biais dans les estimations $\hat{\beta}$ de β
  - ☐ en accroissement de $V(\hat{\beta})$
- Si j'ai un très grand nombre de variables (10 000) et peu d'individus (100), je peux réaliser:
  - ☐ une régression linéaire classique
  - ☐ une sélection de variables backward
  - ☐ une sélection de variable forward
  - ☐ une regression ridge
  - ☐ une régression LASSO

Et parmi elles que je peux réaliser, les classer par ordre de "préférence"

- Si je suis plus intéressé à expliquer les variables importantes dans le modèle plutôt qu'à avoir une bonne qualité de prédiction, je peux réaliser:

- [ ] une regression LASSO
- [ ] une régression Ridge
- [ ] une régression PLS
- [ ] une régression ACR
- [ ] une selection stepwise de variable.