Machine learning I

Eric Matzner-Løber

CEPE

Data Science

Major Influences

Four major influences act today :

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines

Data Science

Major Influences - Tukey (1962)

Four major influences act today :

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines

- He was talking of Data Analysis.
- Data mining, Machine learning, Big Data...

Machine learning et IA

- Data Science : art (ou science) pour extraire la connaissance des données
- Machine learning : création, construction et étude des algorithmes qui apprenent des données et peuvent fournir des prévisions.
- IA : tout outil utilisé par une machine capable de *reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité*

Des problématiques diverses

- Apprentissage supervisé : expliquer/prédire une sortie $y \in \mathcal{Y}$ à partir d'entrées $x \in \mathcal{X}$;
- Apprentissage non supervisé : établir une typologie des observations ;
- Règles d'association : mesurer le lien entre différents produits ;
- Systèmes de recommendation : identifier les produits susceptibles d'intéresser des consommateurs.

Nombreuses applications

finance, économie, marketing, biologie, médecine...

Le problème de la régression

Expliquer une quantité (variable à expliquer) par d'autres (variables **potentiellement** explicatives).

- Objectifs descriptifs : décrire, comprendre la façon dont les variables explicatives agissent sur la quantité à expliquer.
- Objectifs prédictifs : prédire la quantité à expliquer connaissant les variables explicatives

Vocabulaire :

- Lorsque la variable à expliquer est quantitative ($\mathcal{Y} \subset \mathbb{R}^p$), on parle de régression.
- Lorsqu'elle est qualitative (Card($\mathcal{Y}$) fini), de discrimination ou classification supervisée.

Prévision de pics d'ozone

- On a mesuré pendant la concentration maximale quotidienne en ozone
- On dispose également d'autres variables météorologiques (température, nébulosité, vent...).

Individu	O3	T12	Vx	Ne12
1	63.6	13.4	9.35	7
2	89.6	15	5.4	4
3	79	7.9	19.3	8
4	81.2	13.1	12.6	7

Question : peut-on **prédire** la concentration maximale en ozone du lendemain à partir des prévisions météorologiques ?

Détection de spam

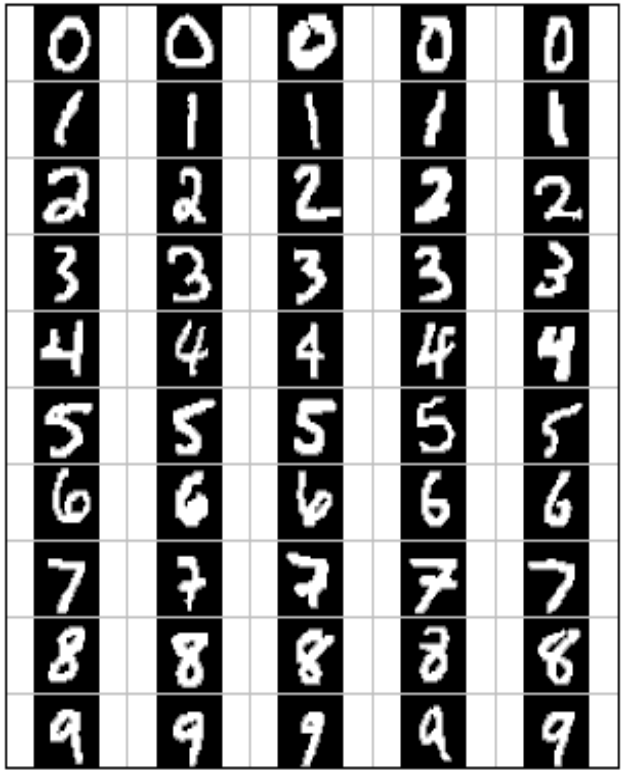
- Sur 4 601 mails, on a pu identifier 1813 spams.
- On a également mesuré sur chacun de ces mails la présence ou absence de 57 mots.

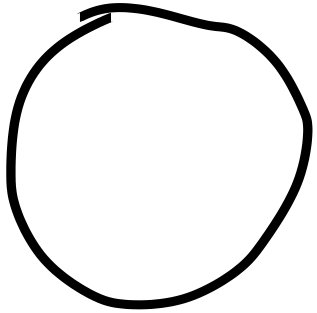
```
> spam[1:5,c(1:8,58)]
  make address  all num3d  our over remove internet type
1 0.00    0.64 0.64     0 0.32 0.00   0.00     0.00 spam
2 0.21    0.28 0.50     0 0.14 0.28   0.21     0.07 spam
3 0.06    0.00 0.71     0 1.23 0.19   0.19     0.12 spam
4 0.00    0.00 0.00     0 0.63 0.00   0.31     0.63 spam
5 0.00    0.00 0.00     0 0.63 0.00   0.31     0.63 spam
```

Question : peut-on construire à partir de ces données une méthode de détection automatique de spam ?

Reconnaissance de l'écriture

Comprendre et apprendre un comportement à partir d'exemples.





Qu'est-ce qui est écrit ? 0, 1, 2... ?

Exemples

La plupart des problèmes présentés précédemment peuvent être appréhendés dans un contexte d'apprentissage supervisé : on cherche à expliquer une sortie y par des entrées x :

y_i	x_i	
C. en O_3	données météo.	Régression
Spam	présence/absence de mots	Discri.
Chiffre	image	Discri.

Remarques

- Les sorties y_i sont représentées par une variable qualitative (discrimination) ou quantitative (régression) ;
- La nature des variables associées aux entrées x_i est variée (quanti, quali, fonctionnelle...).

Un début de formalisation mathématique

- $d_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ on cherche à expliquer/prédire les $y_i \in \mathcal{Y}$ à partir des $x_i \in \mathcal{X}$.
- Il s'agit de trouver une fonction de prévision $f : \mathcal{X} \to \mathcal{Y}$ tq

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

- Nécessité de se donner un critère permettant de mesurer la qualité des fonctions f. On utilise une fonction de perte

$$\ell(y, y') = \begin{cases} 0 & \text{si } y = y' \\ \text{positive} & \text{si } y \neq y'. \end{cases}$$

Approche statistique

- Données $d_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ iid de loi P inconnue.
- Prédicteur : une fonction f de $\mathcal{X} \to \mathcal{Y}$ inconnue à estimer appartenant à $\mathcal{F}$
- Cout : $\ell(f(X), Y)$ qui mesure comment $f(X)$ prévoit Y
- Risque : $\mathcal{R}(f) = \mathbb{E}[\ell(Y, f(X))]$ qui dépend de P

Si on connaissait P, on pourrait trouver la meilleur fonction $f^\star \in \text{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$ appelée oracle.

Objectif
Le travail du data scientist est de trouver un **estimateur** $\hat{f}_n \in \mathcal{F}$ à partir des données tq $\mathcal{R}(\hat{f}_n) \approx \mathcal{R}(f^\star)$.

Choix de la fonction de perte

- Le cadre mathématique développé précédemment sous-entend qu'une fonction est performante (voire optimale) vis à vis d'un **critère** (représenté par la fonction de perte ℓ)).
- Un algorithme de prévision performant pour un critère ne sera pas forcément performant pour un autre.

Conséquence pratique

Avant de s'attacher à construire un algorithme de prévision, il est **capital** de savoir **mesurer la performance** de l'algorithme.

Régression/discrimination

Régression

- La perte est $\ell(y, y') = (y - y')^2$
- Le risque vaut $\mathcal{R}(m) = \mathbb{E}((Y - m(X)^2)$
- Le champion (appelé fonction de régression) est $m^\star(x) = \mathbb{E}[Y|X = x]$

Discrimination

- La perte est $\ell(y, y') = 1_{\{y \neq y'\}}$
- Le risque vaut $\mathcal{R}(m) = \mathbf{P}(g(X) \neq Y)$.
- Le champion, appelé règle de Bayes est

$$g^\star(x) = \begin{cases} -1 & \text{si } \mathbb{P}(Y = -1|X = x) \geq \mathbb{P}(Y = 1|X = x) \\ 1 & \text{sinon,} \end{cases}$$

Machine learning/Statistique

Construire une règle $\hat{f}_n \in \mathcal{F}$ à partir des données tq le risque $R(\hat{f}_n)$ soit petit en moyenne ou avec une forte probabilité respectivement à l'échantillon.

- On resteint la classe $\mathcal{F}$
- On remplace le risque moyen par le risque empirique

$$\hat{f} = \underset{f \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

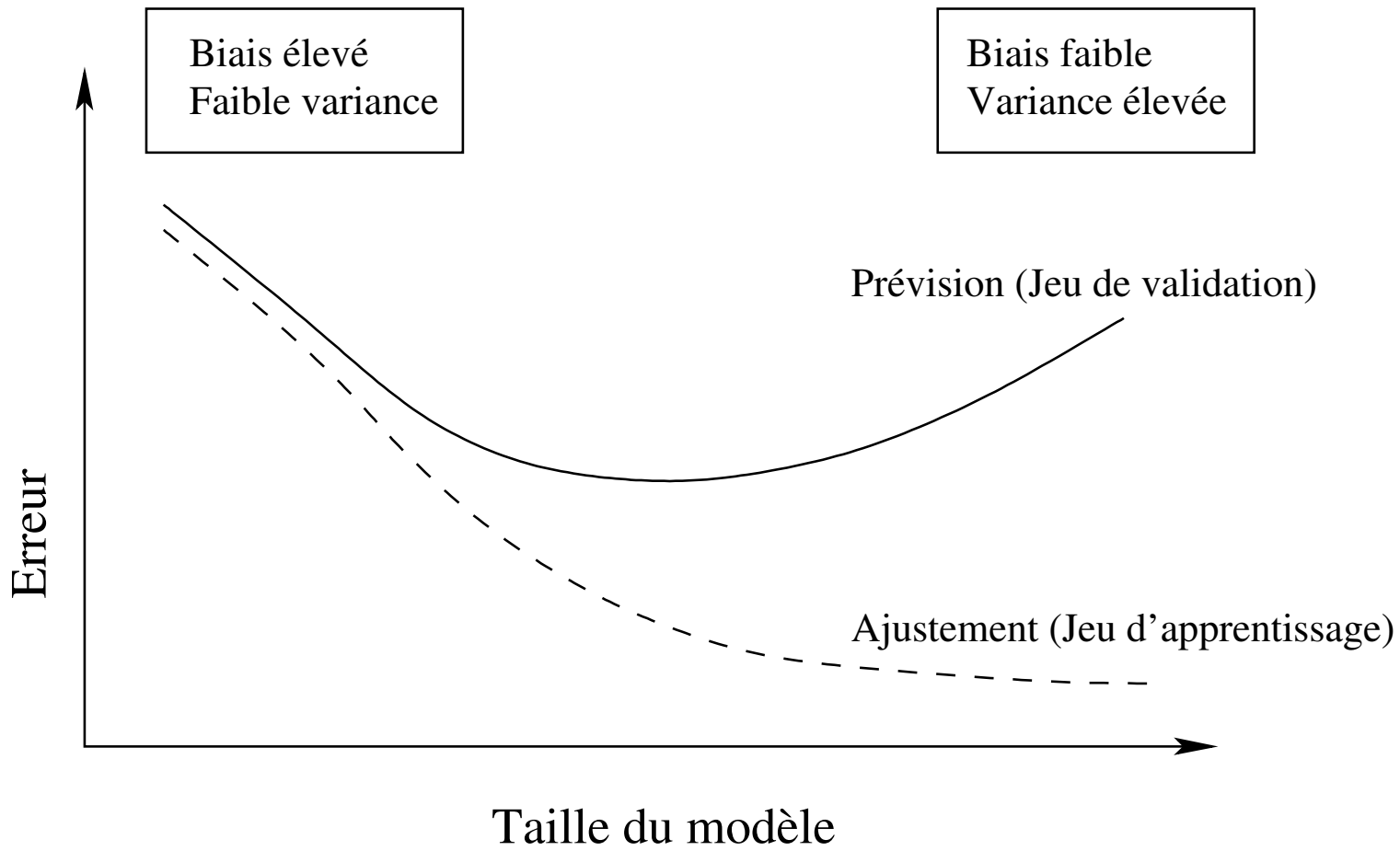
Biais/Variance

- Contexte général $\mathcal{F}$ et $f^* = \text{argmin}_{f \in \mathcal{F}} R(f)$
- Restriction $\mathcal{S} \in \mathcal{F}$, cible $f^*_{\mathcal{S}}$
  or on estime dans $\mathcal{S}$ et on obtient après minimisation $\hat{f}_{\mathcal{S}}$.

$$R(\hat{f}_{\mathcal{S}}) - R(f^*) = \underbrace{R(f^*_{\mathcal{S}}) - R(f^*)}_{\text{erreur d'approximation}} + \underbrace{R(\hat{f}_{\mathcal{S}}) - R(f^*_{\mathcal{S}})}_{\text{erreur d'estimation}}$$

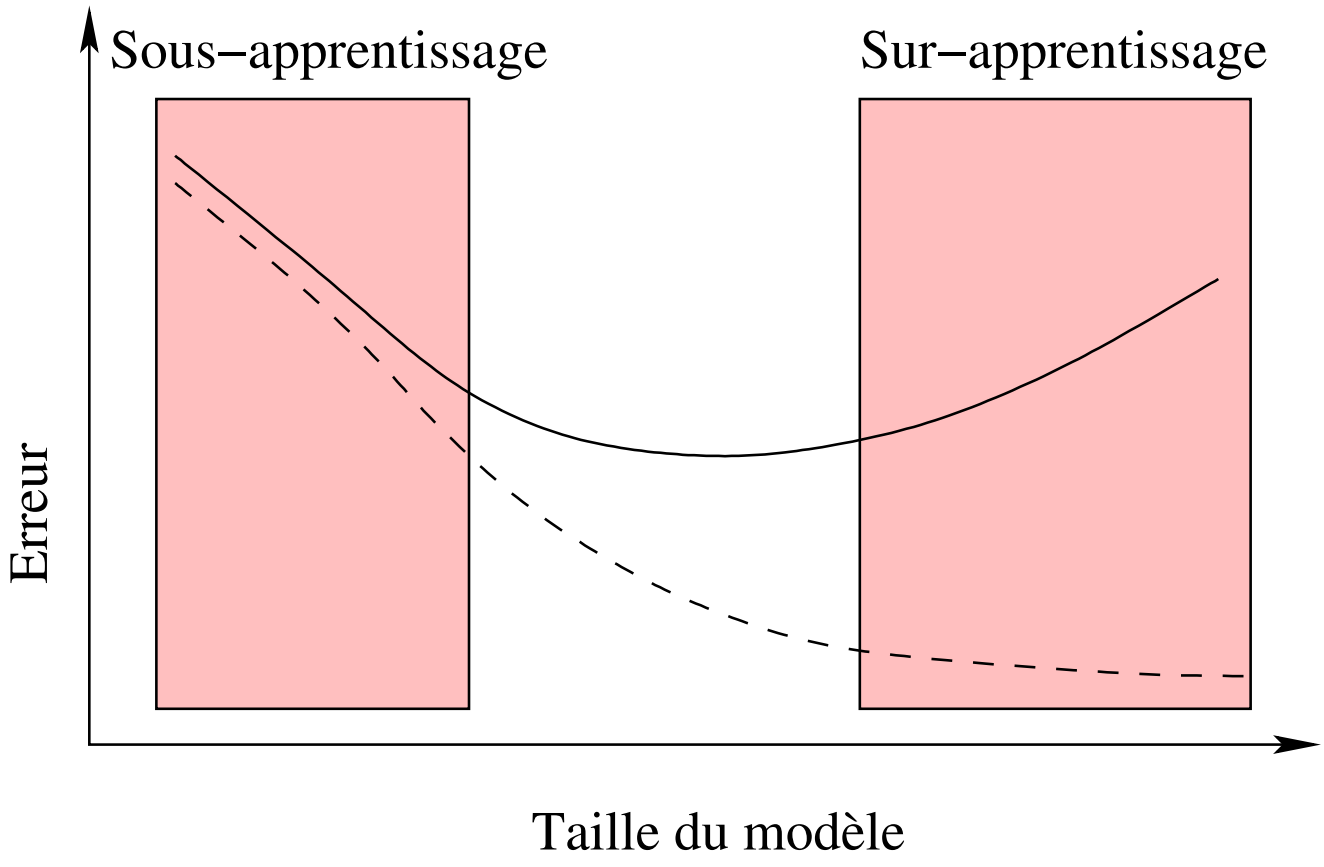
- On apprendra facilement d'un modèle peu complexe mais l'erreur d'approximation (le biais) sera forte.
- On apprendra difficilement d'un modèle très complexe mais l'erreur d'estimation (la variance) sera forte.

Taille de modèle



Idem erreur d'estimation (variance) / erreur d'approximation (biais).

Taille de modèle



Idem erreur d'estimation (variance) / erreur d'approximation (biais).

Surapprentissage

Il est possible d'avoir une erreur d'ajustement faible, on évoquera le sur-apprentissage.
Pour avoir une idée de l'erreur de prévision, il faut avoir un jeu de validation. Si ce jeu n'existe pas, il faut le créer et donc diminuer les données d'apprentissage !

Validation Croisée

- **Idée simple :** utilisation d'un second jeu de données pour calculer l'erreur !
- Suffisant pour éviter le sur-apprentissage !

Cross Validation

- Utilisation de $\frac{V-1}{V}n$ obs pour apprendre et $\frac{1}{V}n$ pour tester.
- En général $V = 5$ ou $V = 10$.

Démarche

- Formalisation du problème
- Recueil et importation des données
- Nettoyage des données et premières analyses
- Découpages et début de modélisation
- Choix des méthodes et modèles, test
- Estimation finale avec toutes les données

Régression linéaire simple

Problème à résoudre

Nous souhaitons expliquer la hauteur d'un eucalyptus par son diamètre afin d'évaluer la hauteur des arbres d'une plantation.

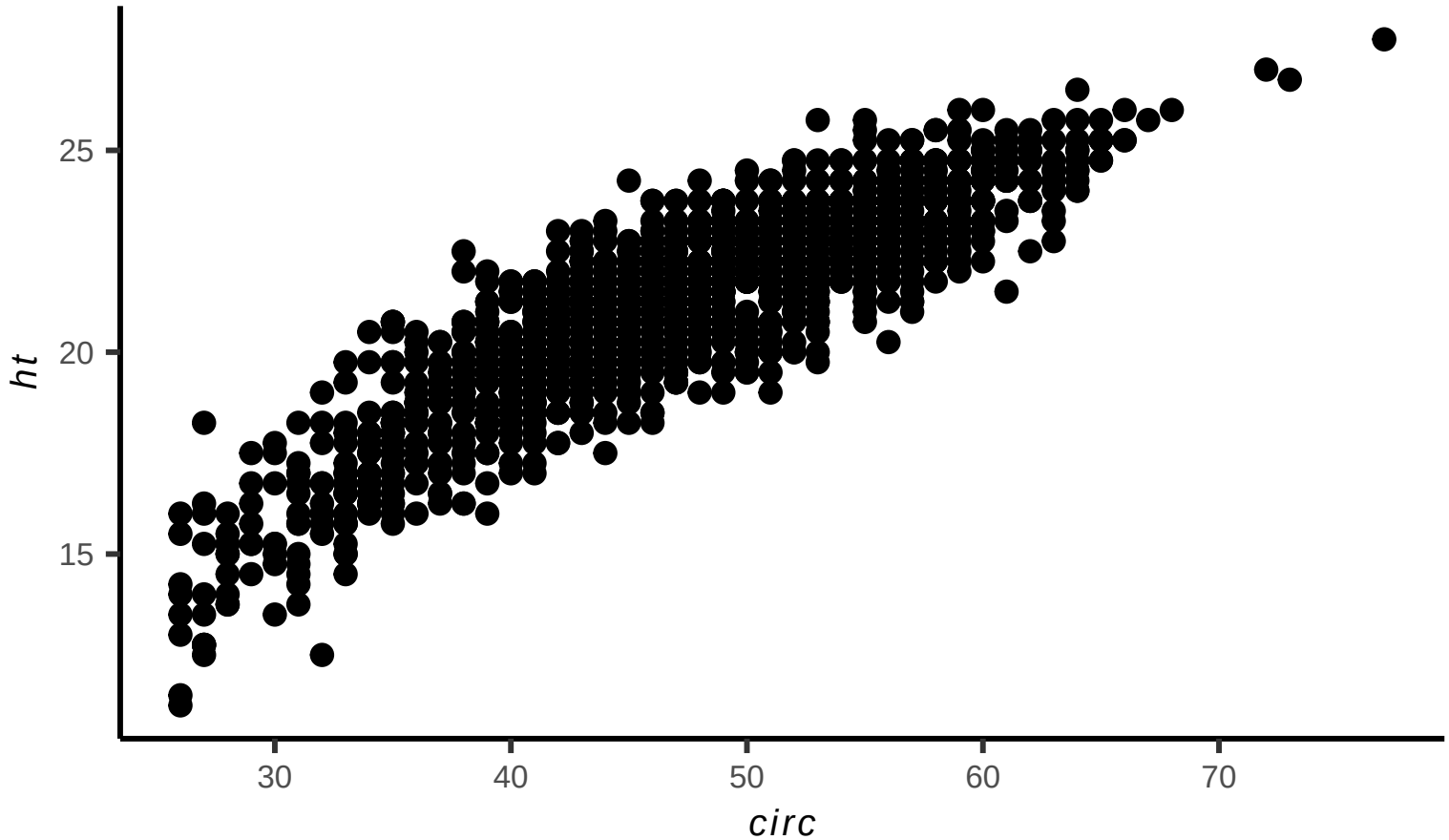
Et il est bien plus facile de mesurer le diamètre que la hauteur !

Mais nous avons quand même réussi à mesurer 1429 arbres

Les données

Individu	ht	circ
1	18.25	36
2	19.75	42
3	16.50	33
4	18.25	39
5	19.50	43
6	16.25	34
7	17.25	37
8	19.00	41
⋮	⋮	⋮

Au total $n = 1429$ mesures que l'on peut représenter



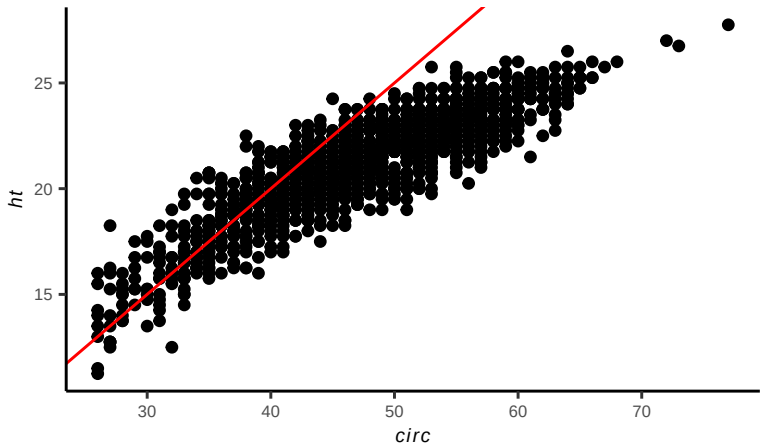
Représentation des $n = 1429$ mesures.

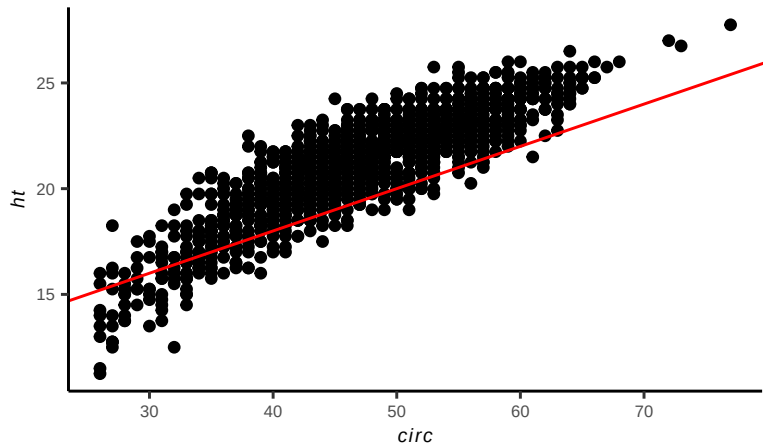
Objectif

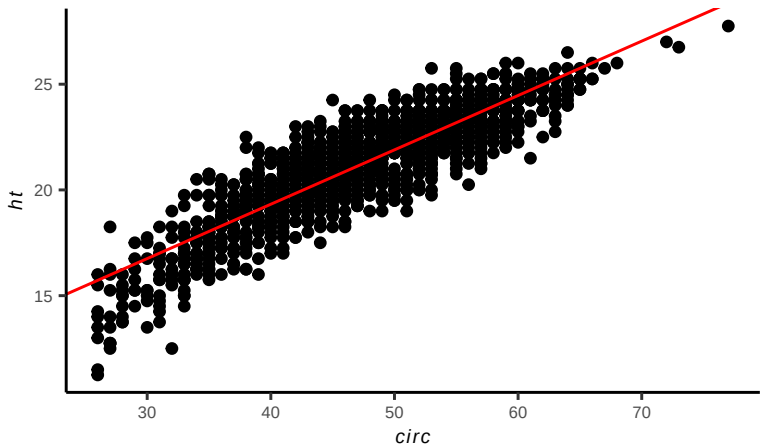
Nous allons chercher une fonction f telle que

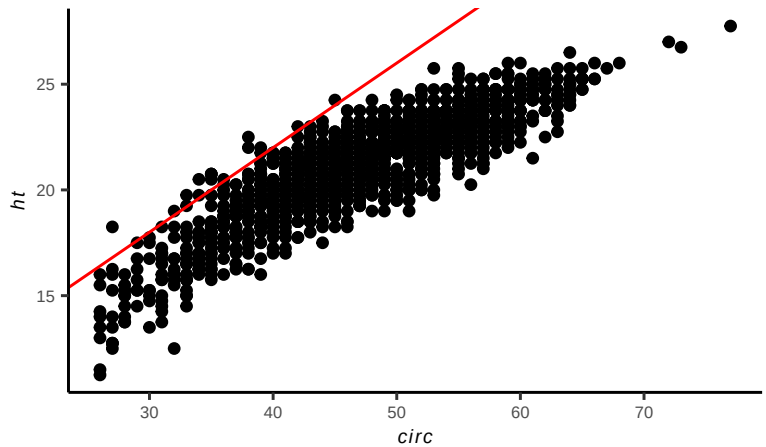
$$\texttt{ht}_i \approx \beta_1 + \beta_2 \texttt{circ}_i.$$

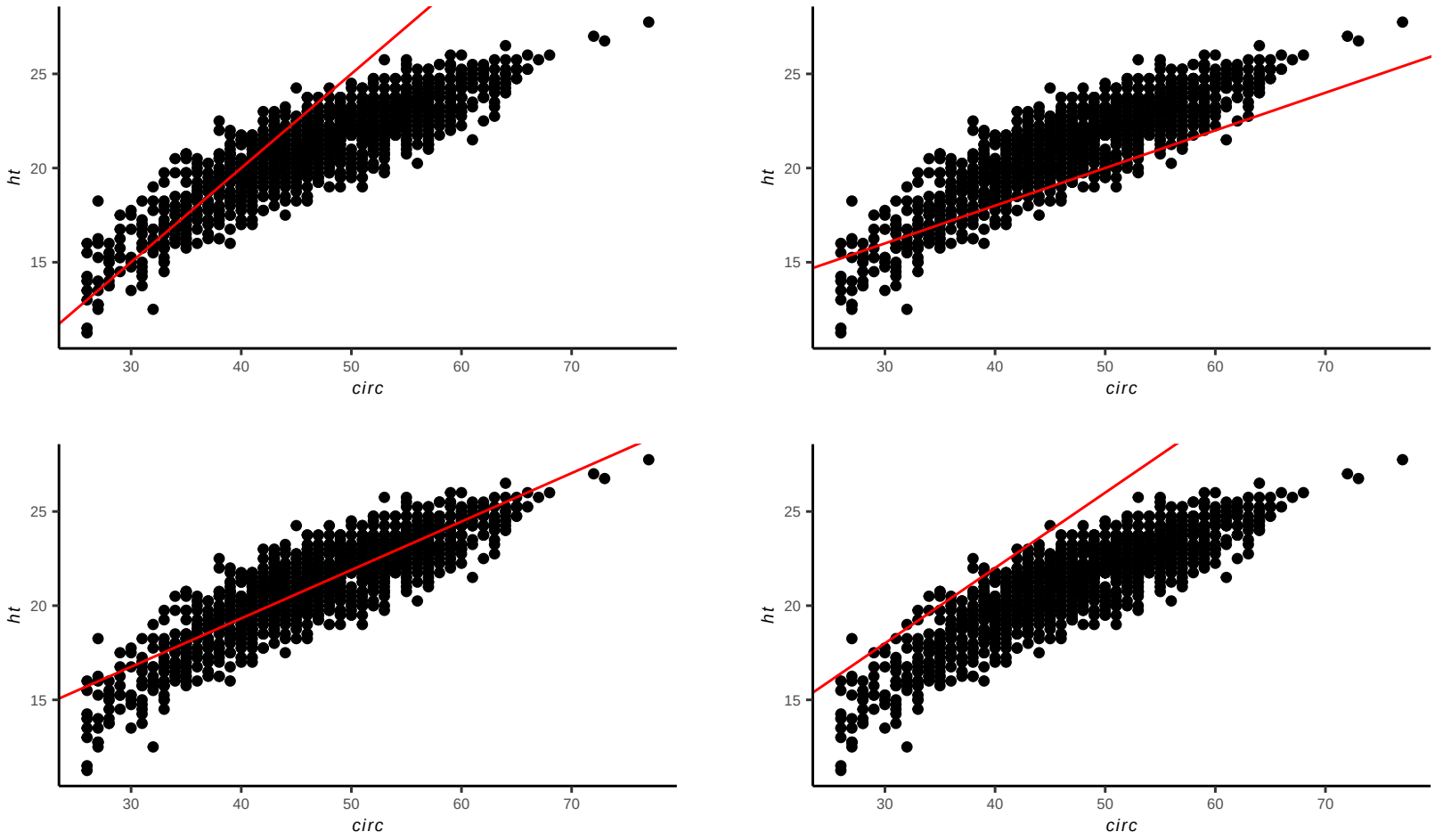
Ce modèle est bien évidemment une approximation de la réalité mais ce modèle imparfait peut être quand même utile.
Nous cherchons donc une droite que passe dans le nuage de points et il en existe une infinité.


ht
25
20
15
30
40
50
60
70
circ


ht
25
20
15
30
40
50
60
70
circ


ht
25
20
15
30
40
50
60
70
circ


ht
25
20
15
30
40
50
60
70
circ



Laquelle choisissez-vous ? Pourquoi ?

Estimateurs des moindres carrés

Il faut choisir un critère appelé fonction de coût.
Nous choisissons le coût quadratique : $l(u) = u^2$.

Nous souhaitons donc minimiser

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

c'est-à-dire que nous souhaitons éviter d'avoir des très fortes erreurs puisqu'elles sont élevées au carré, critère proposé par **Legendre en 1805** car les estimateurs admettent une forme explicite.

NB : Gauss a affimé qu'il utilisait cette méthode depuis 1795 !

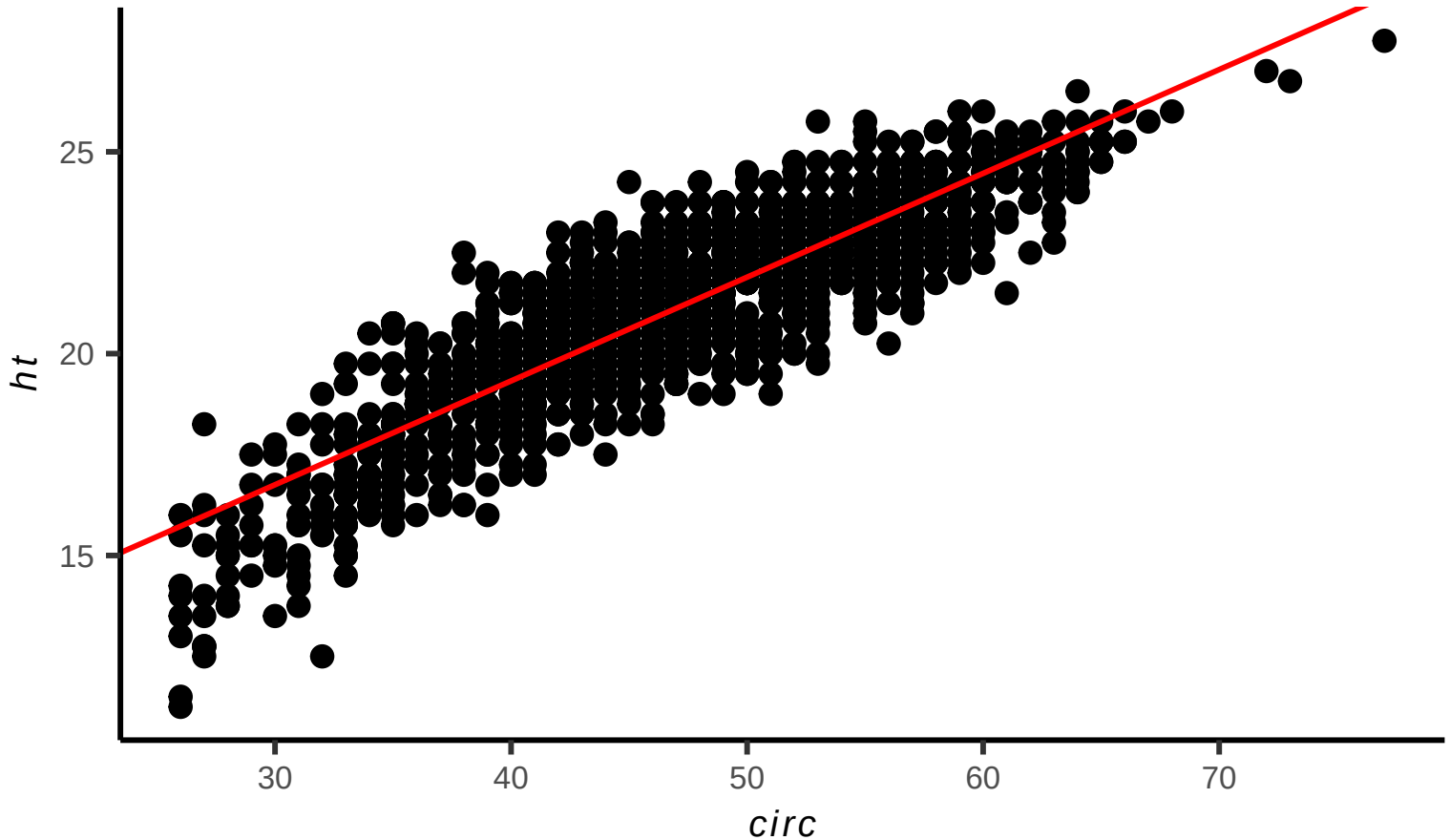
Estimateurs des moindres carrés

Les minimiseurs de $S(\beta_1, \beta_2)$ notés $\text{argmin}_{\beta_1,\beta_2 \in \mathbb{R}^2}(\beta_1, \beta_2)$ sont obtenus en annulant les dérivées de S par rapport à β_1 et β_2 (équations normales) et nous obtenons

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_2 \bar{x}. \end{aligned}$$

avec $\bar{x} = \sum x_i / n$.

Les estimateurs obtenus dépendent des données utilisées et des données différentes donneront des estimateurs différents.



La droite des moindres carrés admet pour équation

$$\widehat{\texttt{ht}} = 9.0375 + 0.2571\texttt{circ}$$

Terminologie

- On **reprend** une circonférence x_i dont le couple (x_i, y_i) a servi à estimer β et on calcule

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

on obtient un $\hat{y}_i$ **estimé** et l'erreur d'estimation est $|y_i - \hat{y}_i|$

- On **mesure un nouveau diamètre** x^* on calcule alors son y^* correspondant

$$\hat{y}^* = \hat{\beta}_1 + \hat{\beta}_2 x^*$$

on obtient un $\hat{y}^*$ **prévu** et l'erreur de prévision est $|y^* - \hat{y}^*|$

En moyenne (erreur estimation) $\leq$ (erreur prévision) naturel non ?

Modèlisation statistique

Considérons le modèle suivant

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

avec comme hypothèses

1. x_i (la circonférence `circ`) fixée
2. ε_i (erreur due à l'imprécision modèle & mesure) : aléatoire
3. Y_i (la hauteur `ht`) : aléatoire
4. β_1, β_2 fixes et inconnus
   $\rightarrow$ β_1, β_2 à estimer...

Estimateurs des moindres carrés

Les estimateurs des Moindres Carrés ($\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2$)

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_2 \bar{x}. \end{aligned}$$

sont des variables aléatoires... Ils dépendent des données et des données différentes donneront des estimateurs différents.

Questions

1. Quelles sont les propriétés des estimateurs ?
2. Quelle est la variablité/précision de l'estimation ?
3. Le modèle est-il bon ?
4. ...

TP

Utiliser les données `eucalyptus.txt` pour retrouver les résultats du cours.

Régression linéaire multiple

Problème à résoudre

La prévision de la qualité de l'air est une problématique importante. Être capable de l'anticiper devrait permettre d'ajuster les politiques publiques afin de prévenir de possibles malades.
Un programme de récolte de données permettant de caractériser la pollution de l'air a été mis en place et il a été mesuré à un point donnée, la concentration d'ozone, la température, la nébulosité, la vitesse et la direction du vent à midi.

Nous souhaitons expliquer la concentration d'ozone (O3) d'un jour donnée et avons récolté 50 journées de mesures.

Les données

Individu	O3	T12	Vx	Ne12
1	63.6	13.4	9.35	7
2	89.6	15	5.4	4
3	79	7.9	19.3	8
4	81.2	13.1	12.6	7
5	88	14.1	-20.3	6

Table – 5 données journalières.

Le vent est mesuré en degré (direction) et m/s (vitesse). Nous avons créé la variable (Vx) projection du vent sur l'axe est-ouest. Au total $n = 50$ impossible à représenter.

Nous allons chercher une fonction f telle que

$$\texttt{O3}_i \approx f(\texttt{T12}_i, \texttt{Vx}_i, \texttt{Ne12}_i).$$

Ce modèle est bien évidemment une approximation de la réalité mais ce modèle imparfait peut être quand même utile.

Le problème mathématique à résoudre est

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_{i1}, \cdots, x_{ip})),$$

Nous choisissons

1. comme fonction de coût le coût quadratique $l(.) = ()^2$

Le problème mathématique à résoudre est

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_{i1}, \cdots, x_{ip})),$$

Nous choisissons

1. comme fonction de coût le coût quadratique $l(.) = ()^2$
2. pour $\mathcal{G}$ nous commençons par les fonctions linéaires :
   $\mathcal{G} = \left\{ f : f(x_1, \cdots, x_p) = \sum_{j=1}^{p} \beta_j x_j \quad \beta_j \in \mathbb{R} \right\}$

Si nous créons la variable $X_{p+1} = \mathtt{T12}^2$, cela reste un modèle linéaire. Ainsi utiliser des polynômes ne change pas la nature de la solution.

Nous souhaitons trouver les valeurs qui minimisent

$$\hat{\beta} = \underset{\beta_1,\cdots,\beta_p}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Un peu de géométrie

Nous avons mesuré n valeurs y_i que nous mettons dans un vecteur de longueur n, noté Y. Ce vecteur est un vecteur de $\mathbb{R}^n$.
Faisons la même chose avec les variables explicatives :

- la constante $\mathbb{1}_n$ dans X_1,
- la température dans X_2,
- la nébulosité dans X_3,
- le vent dans X_4

Concaténons ces vecteurs dans $X = (X_1|X_2|\ldots|X_4)$ de taille $n \times 4$ appelée matrice du plan d'expérience.

Interprétation géométrique

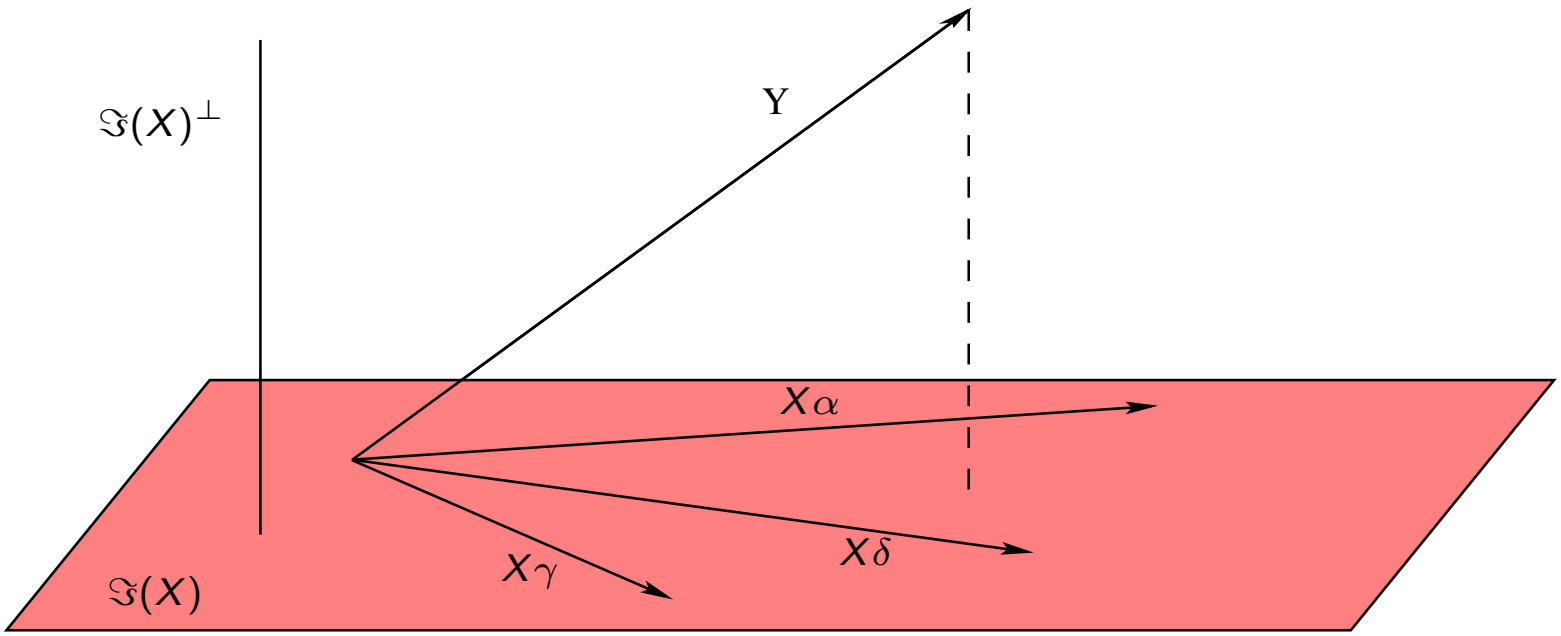


Figure – Représentation dans l'espace des variables.

Ces 4 vecteurs de $\mathbb{R}^n$ engendrent un espace de dimension 4 au maximum noté $\Im(X)$.

Nous cherchons à minimiser

$$\hat{\beta} = \underset{\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

Le minimum est le projeté orthogonal de Y sur $\Im(X)$ le sous-espace engendré par les colonnes de X.

Interprétation géométrique

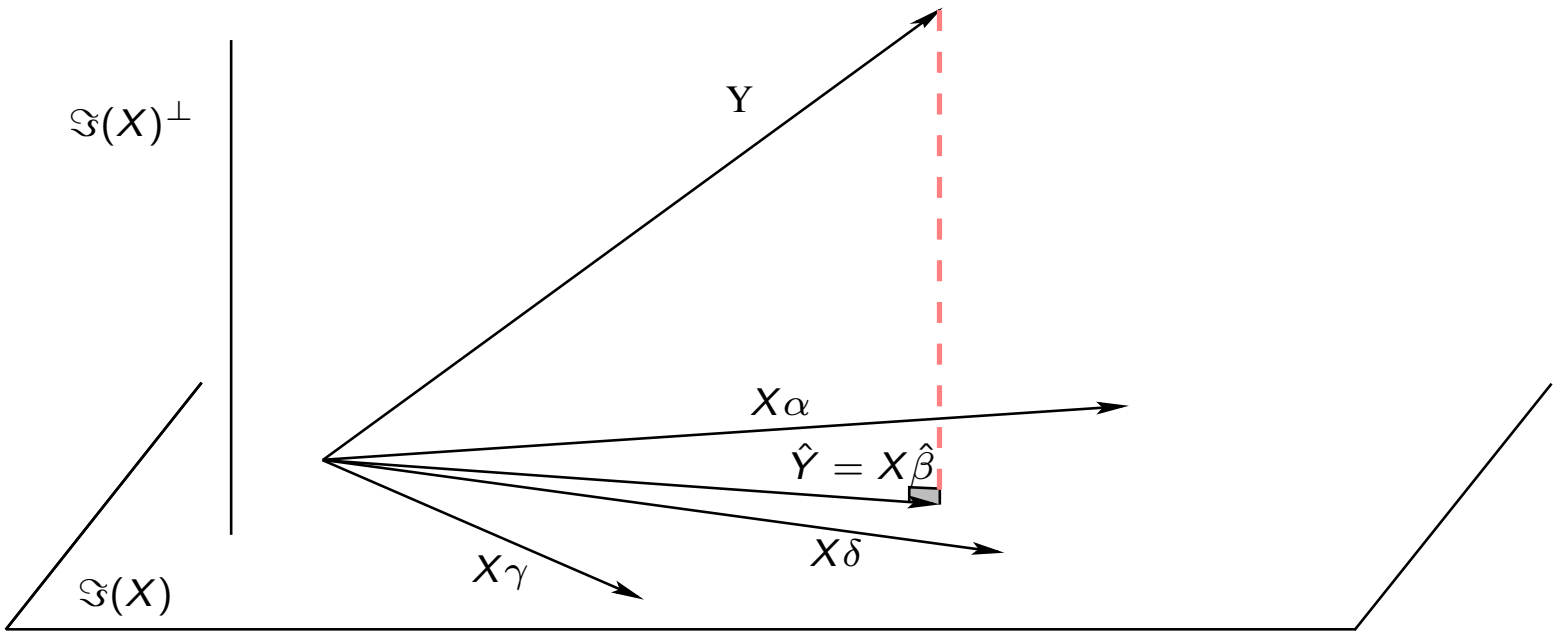


Figure – Représentation dans l'espace des variables.

Ce projeté est unique et son écriture aussi si X est de plein rang.

Est il facile de trouver ce minimum ?

Expression

Si X est de plein rang, l'estimateur des MC $\hat{\beta}$ de β vaut

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

En effet, la matrice de projection orthogonale sur $\Im(X)$ vaut $P_X = X(X'X)^{-1}X'$ et donc comme $P_X Y = X\hat{\beta}$.

On peut aussi dire que $\forall \alpha \in \mathbb{R}^p$, $X\alpha \in \Im(X)$

or $Y - X\hat{\beta} = Y - P_X Y = P_{X^\perp} Y \in \Im(X)^\perp$

donc pour $\forall \alpha \in \mathbb{R}^p$, on a

$$\begin{aligned} < X\alpha, Y - X\hat{\beta} > &= 0. \\ \alpha' X' Y &= \alpha' X X' \hat{\beta}. \end{aligned}$$

Comme cela est vrai pour $\forall \alpha \in \mathbb{R}^p$

$$(XX')^{-1} X' Y = \hat{\beta}.$$

Modélisation statistique

Considérons le modèle suivant

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

où

- les x_{ij} sont des nombres connus, non aléatoires (en général on connaît les variables explicatives : superficie d'un logement, revenu d'un individu...)
- les paramètres à estimer β_j du modèle sont inconnus non aléatoires ; Vecteur de $\mathbb{R}^p$ de coordonnées $(\beta_1, \ldots, \beta_p)$
- les ε_i sont des variables aléatoires inconnues.

Format matriciel

Considérons le modèle suivant

$$Y_{n\times 1} = X_{n\times p}\ \beta_{p\times 1} + \varepsilon_{n\times 1}.$$

- Y vecteur aléatoire de dimension n ;
- X matrice de taille $n \times p$ connue (fixe), appelée matrice du plan d'expérience, X est la concaténation des p variables X_j : $X = (X_1|X_2|\ldots|X_p)$. En général, $X_1 = \mathbb{1}_n$ et β_1 représente l'ordonnée à l'origine (`intercept` en anglais) ;
- β vecteur des paramètres inconnus non aléatoire ($\in \mathbb{R}^p$) ;
- ε vecteur aléatoire centré, de dimension n, des erreurs.

Estimation des MC

On appelle estimateur des moindres carrés (noté MC) $\hat{\beta}$ de β :

$$\hat{\beta} = \underset{\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

Si l'hypothèse $\mathcal{H}_1$ X est de plein rang est vérifiée, l'estimateur des MC $\hat{\beta}$ de β vaut

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Avec ces données, on obtient

$$\hat{y}_i = 84.55 + 1.31\texttt{T12}_i + 0.48\texttt{Vx}_i - 4.89\texttt{Ne12}_i.$$

- On **reprend** un individu x_i dont le couple (x_i, y_i) a servi à estimer β et on calcule $\hat{y}_i = x_i'\hat{\beta}$ on obtient un $\hat{y}_i$ **estimé** et une erreur d'estimation $|y_i - \hat{y}_i|$.

- On **mesure un nouvel** individu x_* on calcule alors son y^* correspondant $\hat{y}^* = x_*'\hat{\beta}$ on obtient un $\hat{y}^*$ **prévu** et une erreur de prévision $|y^* - \hat{y}^*|$.

En moyenne (erreur estimation) $\leq$ (erreur prévision) naturel non ?

Et quelles sont les propriétés de $\hat{\beta}$?

Propriétés Statistiques

Supposons $\mathcal{H}_2$: $\mathbb{E}(\varepsilon_i) = 0 \qquad \text{Var}(\varepsilon) = \sigma^2 \text{I}_n$
Alors

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}((X'X)^{-1}X'Y) \\
&= \mathbb{E}((X'X)^{-1}X'X\beta + X(X'X)^{-1}X'\varepsilon) \\
&= \beta + X(X'X)^{-1}X'\mathbb{E}(\varepsilon) = \beta
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \text{Var}((X'X)^{-1}X'Y) \\
&= (X'X)^{-1}X'\,\text{Var}(Y)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

Mais on ne connaît pas σ^2 !

Alors comment fait on pour avoir un estimateur de la variance de $\hat{\beta}$?

- soit on tire des échantillons de données dans les données initiales et pour chaque tirage on estime $\hat{\beta}$ ce qui donnera la distribution empirique.
- soit on estime σ^2

Résidus et variance

Les résidus sont définis par $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Pour estimer la variance de $\hat{\varepsilon}$ il faut évaluer

$$\sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$$

$\hat{\varepsilon} \in \Im(X)^{perp}$ donc si la constante $\mathbb{1}_n \in \Im(X)$ on a $< \mathbb{1}_n, \hat{\varepsilon} >= 0$ et donc $\sum \hat{\varepsilon}_i = 0$ donc on a l'estimateur suivant

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

On divise par $n - p$ pour avoir un estimateur sans biais et on a un estimateur de la variance de $\hat{\beta}$.

Gauss Markov

Parmi tous les estimateurs **linéaires** et **sans biais** l'estimateur $\hat{\beta}$ est de variance minimale.

Cela sous entend qu'il y aura des estimateurs avec des variances plus petites (estimateurs biaisés, estimateurs non linéaires....)

Est ce que le modèle (qui est faux) est utile ?

TP

Utiliser les données `ozone.txt` pour retrouver les résultats du cours.

Analyse de la modélisation

Nous avons supposé le modèle $Y = X\beta + \varepsilon$.

Sous l'hypothèse $\mathcal{H}_1$ ($\text{rang}(X) = p$), on obtient $\hat{\beta} = (X'X)^{-1}X'Y$.

Sous $\mathcal{H}_2$: $\mathbb{E}(\varepsilon_i) = 0 \qquad \text{Var}(\varepsilon) = \sigma^2 \text{I}_n$

L'estimateur des MC a les propriétés suivantes

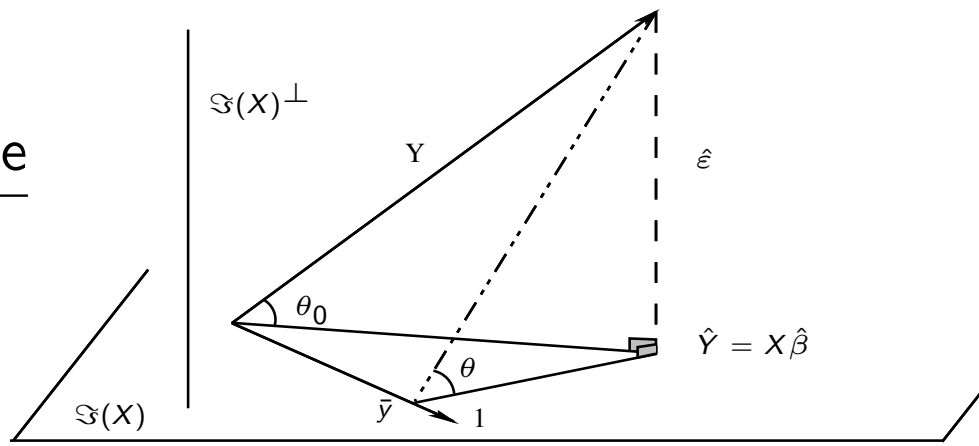
$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \beta \\ \text{Var}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\ \hat{\sigma}^2 &= \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \end{aligned}$$

R^2

Un critère utilisé pour analyser la qualité d'une régression est le

$$R^2 = \frac{\text{V. expliquée par le modèle}}{\text{Variation totale}}$$

$$= \frac{\|\hat{Y} - \bar{y}1\|^2}{\|Y - \bar{y}1\|^2} = \cos^2\theta.$$



Le R^2 augmente avec $\dim(\Im(X))$, pas très intéressant !

Pour faciliter les calculs remarquons que

$$\hat{\varepsilon} \quad = \quad Y - \hat{Y} = (I - P_x)Y = P_{X^\perp} Y = P_{X^\perp}\varepsilon$$

$$\mathbb{E}(\hat{\varepsilon}) = 0 \quad \text{Var}(\hat{\varepsilon}) = P_{X^\perp}\sigma^2.$$

Ainsi $\hat{\varepsilon}$ n'a pas les mêmes propiétés que ε.

- ε_i homoscédastiques (même variance) et non corrélés
- $\hat{\varepsilon}_i$ hétéroscédastiques et corrélés.

Il va falloir utiliser d'autres résidus si on veut les analyser.
Notons h_{ij} les éléments de la matrice de projection P_X.

Définitions

- résidus $\hat{\varepsilon}_i = y_i - \hat{y}_i$.
- résidus normalisés $r_i = \hat{\varepsilon}_i/(\sigma\sqrt{1-h_{ii}})$
- résidus standardisés $t_i = \hat{\varepsilon}_i/(\hat{\sigma}\sqrt{1-h_{ii}})$
- résidus studentisés par Validation Croisée

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

où $\hat{\sigma}_{(i)}$ signifie que cet estimateur a été obtenu sans l'individu i

```
infl = modele.get_influence()
infl.resid_studentized_external
```

ou **rstudent** dans R.

Valeur aberrante

Si on suppose que les ε suivent une loi normale on peut montrer que les résidus studentisés par Validation Croisée suivent une loi de Student.

On peut donc définir une valeur aberrante.

Valeur aberrante

Si on suppose que les ε suivent une loi normale on peut montrer que les résidus studentisés par Validation Croisée suivent une loi de Student.

On peut donc définir une valeur aberrante. C'est un individu (x_i', y_i) pour lequel la valeur associée à t_i^* est élevée (comparée au seuil donné par la loi du Student) : $|t_i^*| > t_{n-p-1}(1-\alpha/2)$.

Valeur aberrante

C'est un individu (x_i', y_i) pour lequel la valeur associée à t_i^* est élevée (comparée au seuil donné par la loi du Student) : $|t_i^*| > t_{n-p-1}(1-\alpha/2)$.

Attention, il y aura en moyenne $\alpha\%$ des points en dehors de l'intervalle, donc utiliser son bon sens et relier la valeur de t_i^* à sa probabilité et la taille de l'échantillon.

Exemple, $t_i^* = 3$, cela devrait arriver 1 fois sur 1000.
Les points aberrants sont à analyser de façon séquentielle.

Question

Une valeur aberrante est détectée car son y_i et son estimé $\hat{y}_i$ diffèrent mais quand est-il si des individus sont très différents vis à vis de leurs variables explicatives x_i' ?

Matrice de projection

En développant l'écriture des valeurs ajustées, nous obtenons

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

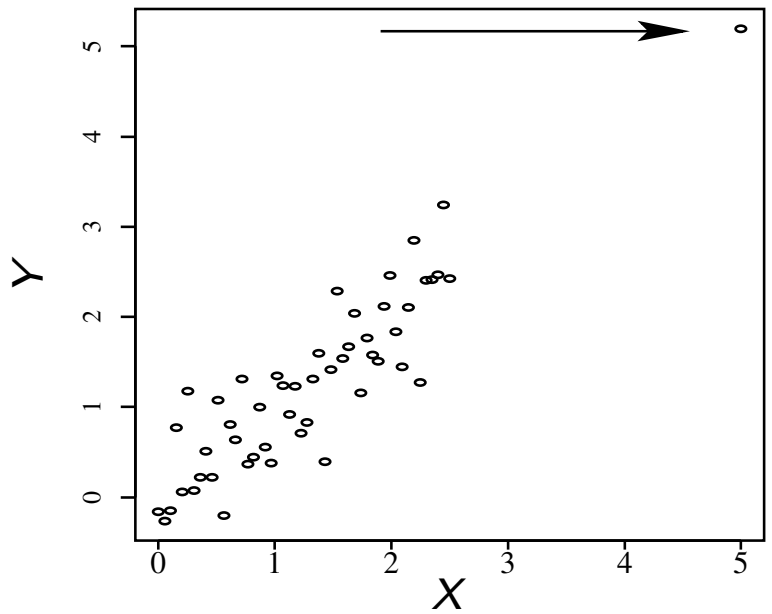
le poids de l'observation i sur sa propre estimation (h_{ii}). Nous avons les cas extrêmes suivants :

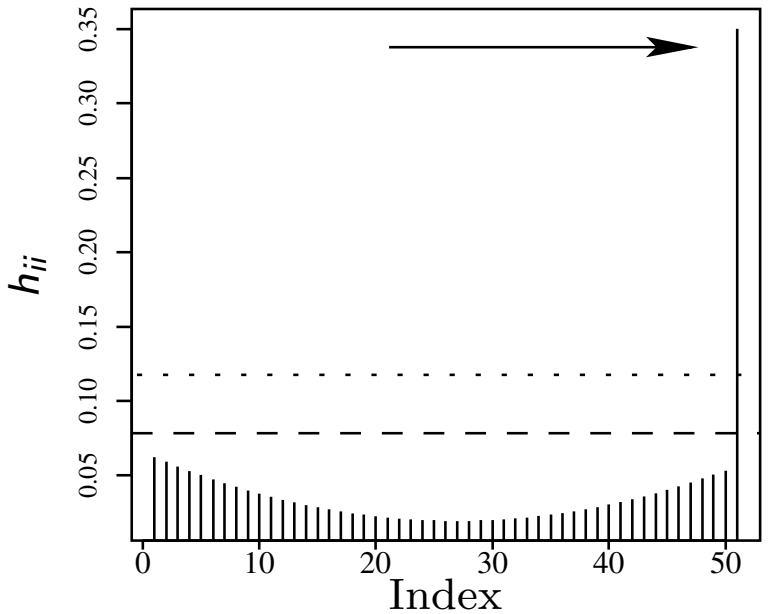
- si $h_{ii} = 1$, $\hat{y}_i$ est entièrement déterminée par y_i car $h_{ij} = 0$ pour tout j ;
- si $h_{ii} = 0$, y_i n'a pas d'influence sur $\hat{y}_i$ (qui vaut alors zéro).

Point levier

Un individu i est dit levier si sa valeur h_{ii} est bien plus grande que les autres valeurs.

Exemple d'un point levier, figuré par la flèche, pour un modèle de régression simple. La ligne en pointillé représente le seuil de $3p/n$ et celle en tiret le seuil de $2p/n$.





Résumé

A la fin de l'étape d'estimation, il est conseillé d'analyser

- les résidus studentisés
- les points leviers

Les résidus studentisés ne doivent pas avoir de structure donc il faut penser à faire des analyses graphiques.

Choix de variables

Est ce que toutes les variables sont utiles à la modélisation ?

$$Y = X_\xi \beta + \varepsilon$$

Problèmes :

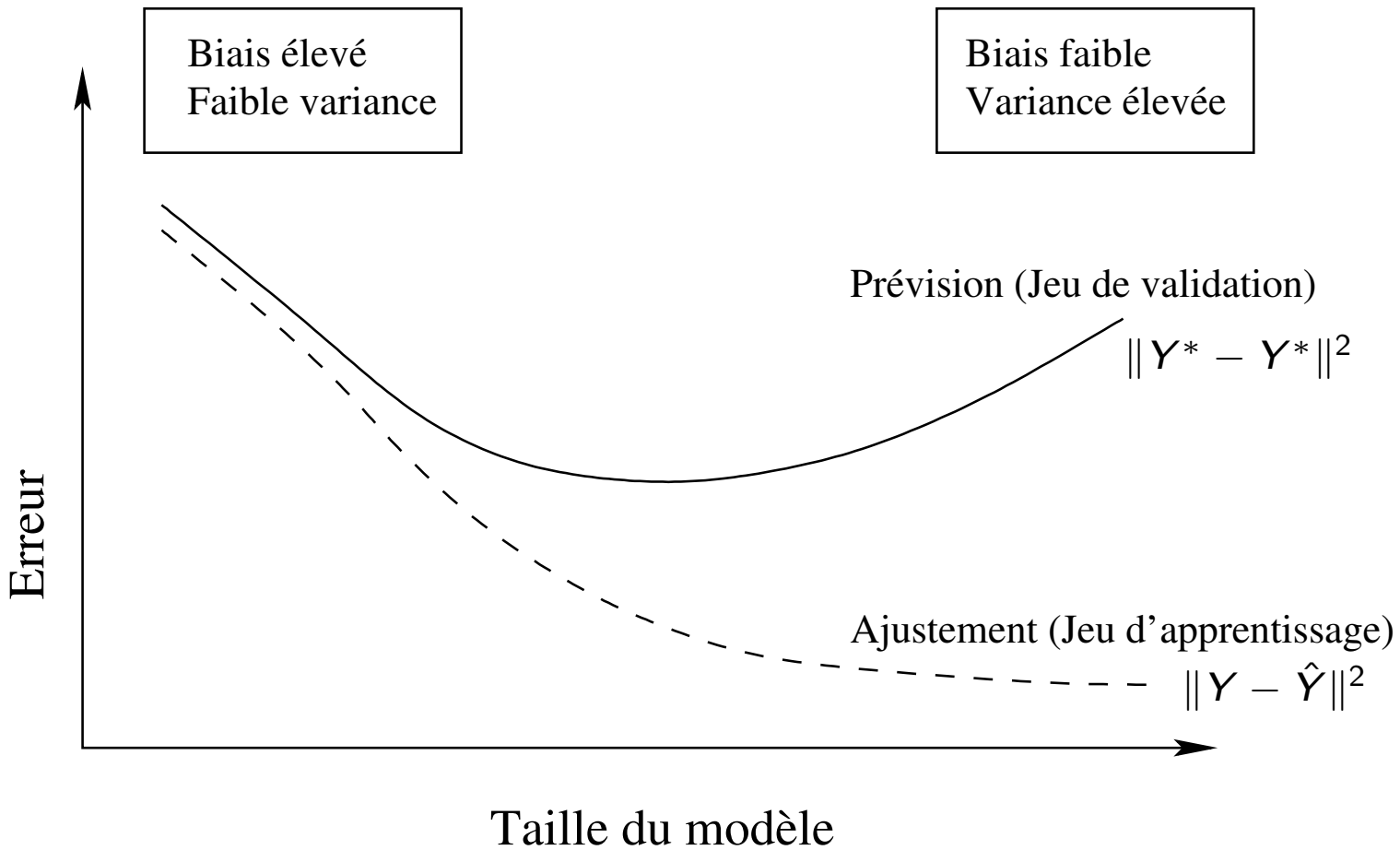
1. Estimation de ξ (les variables à sélectionner)
2. Estimation des coefficients : $\hat{\beta}_{\hat{\xi}}$

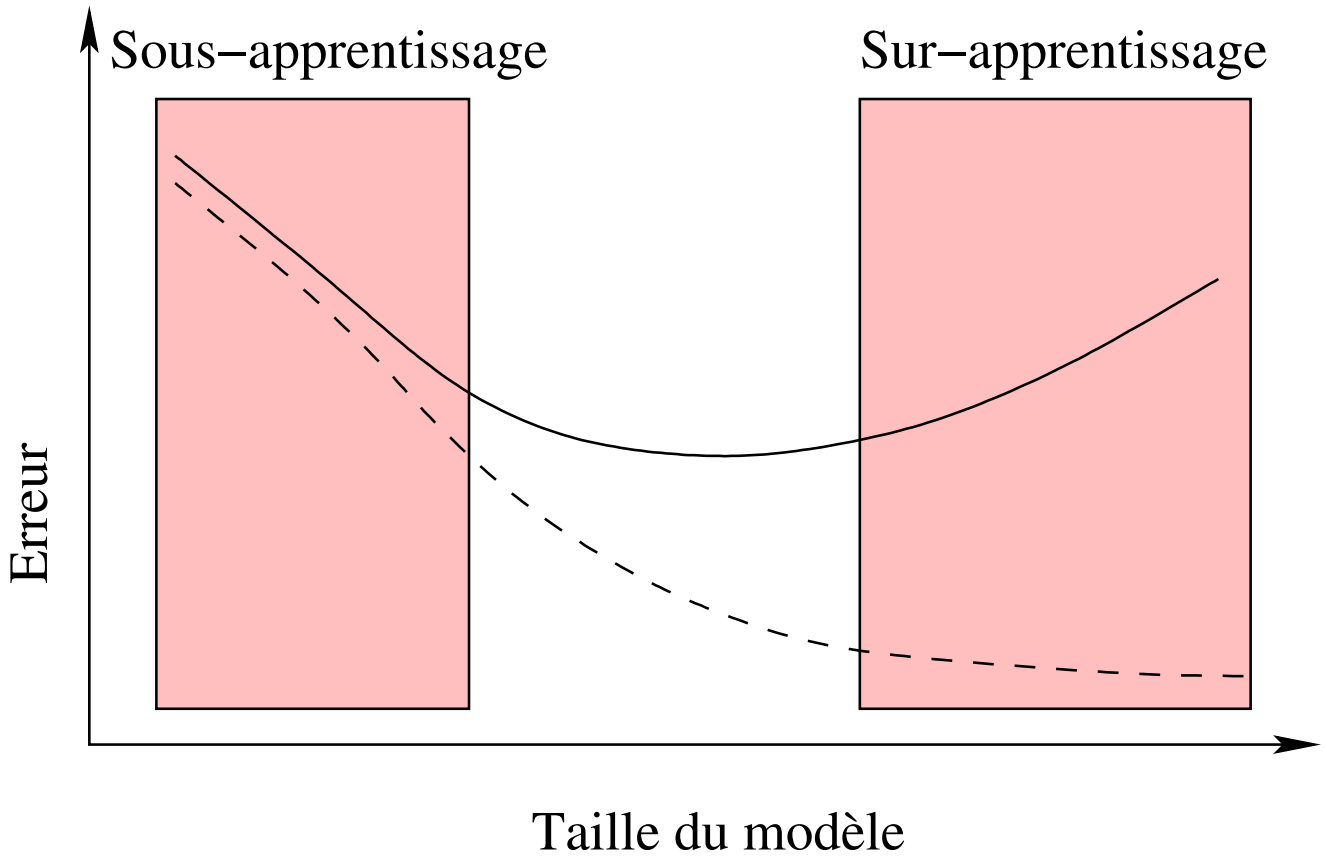
Plus le nombre de variables augmente

- $\hat{\beta}_{\hat{\xi}}$ moins précis (plus de variance)
- On ne rate pas les variables explicatives (moins de biais)

Conclusion : il faut trouver un compromis entre les 2 phénomènes

Taille de modèle





Il est possible d'avoir une erreur d'ajustement faible, on évoquera le sur-apprentissage.
Pour avoir une idée de l'erreur de prévision, il faut avoir un jeu de validation. Si ce jeu n'existe pas, il faut le créer et donc diminuer les données d'apprentissage !

Si les modèles sont emboîtés, nous avons vu la dernière fois le test F qui possède comme loi sous H_0 :

$$F = \frac{\|\hat{Y}_0 - \hat{Y}\|^2/(p - p_0)}{\|Y - \hat{Y}\|^2/(n - p)} \sim \mathcal{F}_{p-p_0,n-p},$$

L'hypothèse H_0 sera repoussée en faveur de H_1 si l'observation de la statistique F est supérieure à $f_{p-p_0,n-p}(1-\alpha)$, la valeur α est le niveau du test.

Mais si

on me demande de choisir entre ces 2 modélisations

1. Modèle 1 :

$$\texttt{O3} = \beta_1 + \beta_2 \texttt{T12} + \varepsilon$$

2. Modèle 2 :

$$\texttt{O3} = \beta_1 + \beta_2 \texttt{Vx} + \beta_3 \texttt{Ne12} + \varepsilon$$

Comment est ce que je fais ?

Je ne peux pas faire un test car les modèles ne sont pas emboîtés (mince !)

Je peux calculer l'erreur de prévision ?

Très bonne idée, mais pour cela il faudrait que j'ai un autre jeu de données que je n'ai pas utilisé pour l'estimation. J'aurais pu mettre des données de côté avant mais maintenant c'est trop tard on y pensera la prochaine fois.

Si je suppose la normalité de ε je peux calculer la vraisemblance de chaque modélisation ?
Oui mais nous avons vu que

$$\text{maximiser } \mathcal{L}(Y, \beta, \sigma^2) \text{ idem minimiser } \|Y - X\beta\|^2$$

cela pourrait favoriser les modélisations avec beaucoup de paramètres.

Il a été proposé de pénaliser les modèles qui ont beaucoup de variables (principe de parcimonie)

Critères de choix AIC et BIC

$$\begin{aligned} AIC &= -2\ln(\mathcal{L}) + 2 \times \text{nb. param.} \\ BIC &= -2\ln(\mathcal{L}) + ln(n) \times \text{nb. param.} \end{aligned}$$

Choix de modèles (en régression) par critère de choix

Idée naïve

1. Estimation de tous les modèles possibles ($2^p - 1$)
2. Calcul du BIC (ou AIC ou...)
3. Sélection du modèle avec meilleur critère

Problème du temps de calcul

- approche exhaustive par algo. adapté
- approche non exhaustive par forward, backward, stepwise

Algorithme Backward

- Calcul du critère pour le modèle à p variables noté M_0.
- Calcul du critère pour les p modèles à $p-1$.
- Choix du meilleur modèle noté M_1
- Si $M_0 < M_-$ on conserve M_0 et on arrête.
- Si $M_1 < M_0$ le modèle M_1 devient M_0 on recommence.

Algorithme Backward

Utiliser la fonction `step` pour effectuer le choix de variables sur les données d'ozone.

Variables qualitatives

Parmi les données potentiellement explicatives, il peut y avoir des variables qualitatives.

- Comment sont-elles codées ?
- Le codage a-t-il un effet sur les résultats ?
- Faut il réduire les variables ?
- ...

Les données

Individu	O3	T12	Ne	Dv
1	63.6	13.4	n	E
2	89.6	15	s	N
3	79	7.9	n	E
4	81.2	13.1	n	N
5	88	14.1	n	O
6	68.4	16.7	n	S

Table – 6 données journalières.

Modélisation naïve

Nous « modélisons » les données par

$$\texttt{O3}_i = \beta_1 + \beta_T \texttt{T12}_i + \beta_{Ne} \texttt{Ne}_i + \beta_{Dv} \texttt{Dv}_i + \varepsilon_i$$

Problème

La matrice du plan d'expérience devrait être :

$$X : \begin{array}{c} \begin{matrix} \mathbb{1} & \texttt{T12} & \texttt{Ne} & \texttt{Dv} \end{matrix} \\ \begin{pmatrix} 1 & 13.4 & n & E \\ 1 & 15 & s & N \\ 1 & 7.9 & n & E \\ 1 & 13.1 & n & N \\ 1 & 14.1 & n & O \\ 1 & 16.7 & n & S \end{pmatrix} \end{array} \tag{1}$$

Codage disjonctif complet

On va utiliser un coefficient par modalité et le modèle devient :

$$\begin{aligned} \texttt{O3}_i &= \beta_1 + \beta_T \texttt{T12}_i + \beta_n \mathbb{1}_{ni} + \beta_s \mathbb{1}_{si} + \\ &\quad \beta_E \mathbb{1}_{Ei} + \beta_N \mathbb{1}_{Ni} + \beta_O \mathbb{1}_{Oi} + \beta_S \mathbb{1}_{Si} + \varepsilon_i \end{aligned}$$

Estimation par MCO classique avec la matrice du plan d'expérience

$$X = \begin{array}{c} \begin{array}{cccccccc} \mathbb{1} & \texttt{T12} & \mathbb{1}_n & \mathbb{1}_s & \mathbb{1}_E & \mathbb{1}_N & \mathbb{1}_O & \mathbb{1}_S \end{array} \\ \left[\begin{array}{cccccccc} 1 & 13.4 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 15 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 7.9 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 13.1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 14.1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 16.7 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

La matrice n'est pas de plein rang

La solution classique est de supprimer une modalité par variable qualitative, en général la première (modalité de référence), cela donne donc

$$X = \begin{array}{c} \begin{array}{cccccc} \mathbb{1} & \texttt{T12} & \mathbb{1}_S & \mathbb{1}_N & \mathbb{1}_O & \mathbb{1}_S \end{array} \\ \left[\begin{array}{cc:cccc} 1 & 13.4 & 0 & 0 & 0 & 0 \\ 1 & 15 & 1 & 1 & 0 & 0 \\ 1 & 7.9 & 0 & 0 & 0 & 0 \\ 1 & 13.1 & 0 & 1 & 0 & 0 \\ 1 & 14.1 & 0 & 0 & 1 & 0 \\ 1 & 16.7 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

Le modèle devient

$$\texttt{O3}_i = \beta_1 + \beta_T \texttt{T12}_i + \beta_s \mathbb{1}_{Si} + \beta_N \mathbb{1}_{Ni} + \beta_O \mathbb{1}_{Oi} + \beta_S \mathbb{1}_{Si} + \varepsilon_i$$

Constante ou intercept

Définition (Constante ou intercept)

La constante est la valeur que prend le modèle quand toutes les variables explicatives valent 0.

Constante dans notre modèle

Modèle (contrainte « traitement » ou « référence »)

$$\texttt{O3}_i = \beta_1 + \beta_T \texttt{T12}_i + \beta_s \mathbb{1}_{Si} + \beta_N \mathbb{1}_{Ni} + \beta_O \mathbb{1}_{Oi} + \beta_S \mathbb{1}_{Si} + \varepsilon_i$$

La constante β_1 est donc la valeur pour

- $\texttt{T12}_i = 0$, $\mathbb{1}_{Si} = 0$, $\mathbb{1}_{Ni} = 0$, $\mathbb{1}_{Oi} = 0$, $\mathbb{1}_{Si} = 0$
- $\texttt{T12}_i = 0$ et pour Ne=nuage et pour Dv=Est

Coefficients

Sens des coefficients dans notre modèle

Modèle

$$\texttt{O3}_i = \beta_1 + \beta_T \texttt{T12}_i + \beta_s \mathbb{1}_{si} + \beta_N \mathbb{1}_{Ni} + \beta_O \mathbb{1}_{Oi} + \beta_S \mathbb{1}_{Si} + \varepsilon_i$$

- β_T effet linéaire de `T12`
- β_s écart quand on passe de `n` (référence) à `s`
- β_O écart quand on passe de `E` (référence) à `O`
- β_N écart quand on passe de `E` (référence) à `N`
- β_S écart quand on passe de `E` (référence) à `S`

Régression Ridge

Objectif

Nous avons supposé que le modèle de régression

$$Y = X\beta + \varepsilon$$

était correct et que la matrice X était de plein rang.

Or si

- $n < p$, le nombre de variables est supérieur au nombre d'observations (grande dimension)
- $n \geq p$ mais $\{X_1, \ldots, X_p\}$ est une famille liée de $\mathbb{R}^n$.

$X'X$ n'est pas de plein rang, $X'X$ n'est pas inversible, la relation donnant $\hat{\beta}$ n'a pas de sens mais le projeté de Y sur $\Im(X)$ existe toujours.

Solution basique

Rendre $X'X$ inversible

en rajoutant des termes sur la diagonale :

$$X'X \to X'X + \lambda \mathrm{I}_p$$

Cela donne l'estimateur ridge (1970) :

$$\hat{\beta}_{\mathrm{ridge}}(\lambda) = (X'X + \lambda \mathrm{I}_p)^{-1} X'Y$$

OK on rajoute λ sur la diagonale pour rendre inversible $X'X$ mais y aurait il une autre explication à ridge ?

Minimisation des MC pénalisés

Imaginons que l'on nous demande pour un λ donné de minimiser les MC pénalisées

$$\|Y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

- Si $\lambda = 0$ on retrouve les estimateurs de MC.
- Si λ grand, on va forcer à avoir $\hat{\beta}$ petit afin que le terme de droite soit petit.

Conclusion : on va rétrécir les estimateurs $\hat{\beta}$ et $\hat{Y}(\lambda)$ se rapprochent de l'origine.

Minimisation des MCO pénalisés

Minimisons pour un λ donné de minimiser les MC pénalisées

$$\|Y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

Ecrivons cela autrement

$$(Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$

Dérivons

$$2(-X')(Y - X\beta) + 2\lambda\beta.$$

Annulons les dérivées

$$\begin{aligned}(X'X + \lambda I)\hat{\beta} &= X'Y \\ \hat{\beta}(\lambda) &= (X'X + \lambda I)^{-1}X'Y\end{aligned}$$

On retrouve la régression ridge.

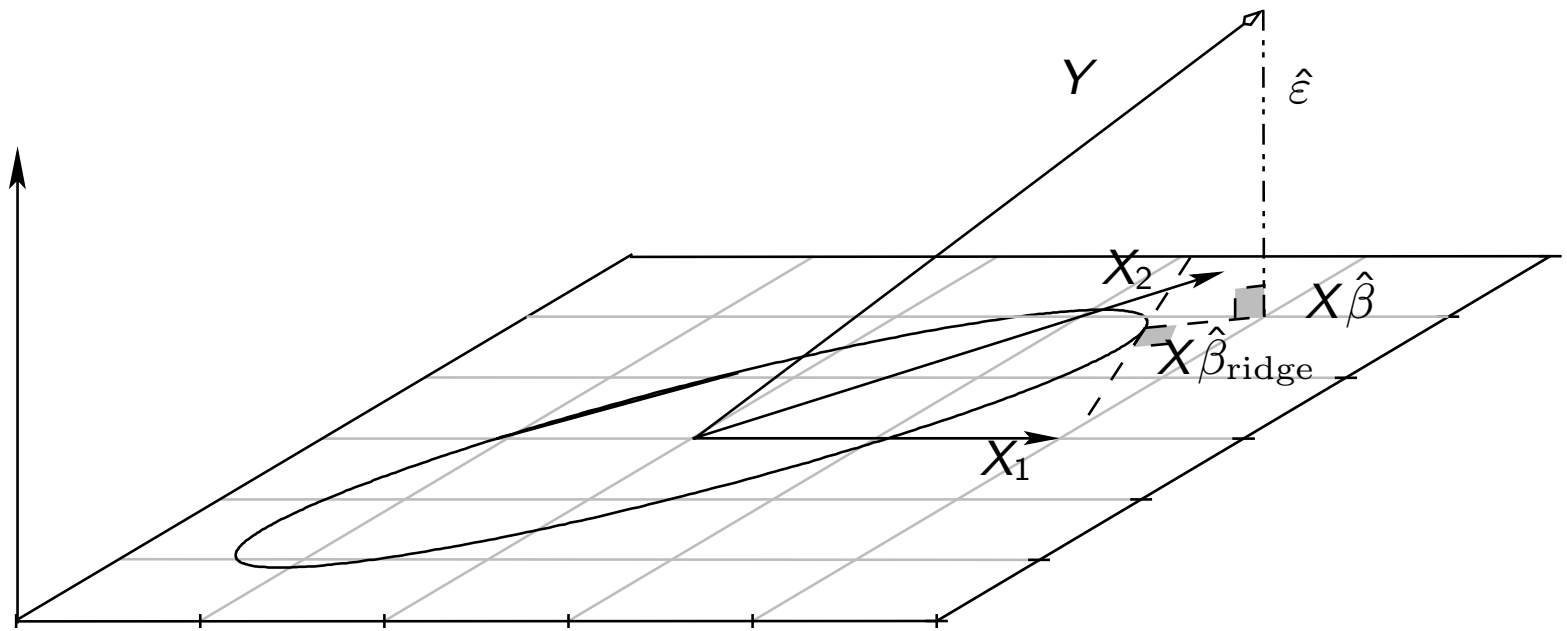
Régression sous contrainte

$$\min_{\beta} \|Y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

ce qui peut aussi s'écrire sous la forme suivante

$$\hat{\beta}(\delta) = \underset{\beta\in\mathbb{R}^p, \|\beta\|_2^2 \leq \delta}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

Et donc à nouveau on rétrécit l'estimateur des MC.


Y
$\hat{\varepsilon}$
X_2
$X\hat{\beta}$
$X\hat{\beta}_{\text{ridge}}$
X_1

Interprétation statistique

$$\begin{aligned} \hat{\beta}_{MC} &= (X'X)^{-1}X'Y \\ \hat{\beta}_{\text{ridge}}(\lambda) &= (X'X + \lambda \mathrm{I}_p)^{-1}X'Y \end{aligned}$$

On peut donc écrire l'un en fonction de l'autre

$$\hat{\beta}_{\text{ridge}}(\lambda) = (X'X + \lambda \mathrm{I}_p)^{-1}(X'X)\hat{\beta}_{MC}$$

Propriétés statistiques

$$\begin{aligned} \mathbb{E}(\hat{\beta}_{MC}) &= \beta \\ \text{Var}(\hat{\beta}_{MC}) &= \sigma^2(X'X)^{-1} \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\hat{\beta}_{\text{ridge}}) &= (X'X + \lambda \text{I})^{-1}(X'X)\beta \\ \text{Var}(\hat{\beta}_{\text{ridge}}) &= \sigma^2(X'X + \lambda \text{I})^{-1}X'X(X'X + \lambda \text{I})^{-1} \end{aligned}$$

Remarque

Pour comparer $EQM = biais^2 + \text{Var}$ ou sa trace.

Qualité de l'estimation EQM

En utilisant $X'X = P\,\mathrm{diag}(\lambda_i)P'$,

on peut montrer que

$$\begin{aligned}\mathrm{trace}(EQM(\hat{\beta}_{MC})) &= 0 + \sigma^2\,\mathrm{trace}(X'X)^{-1} = \sum_{j=1}^{p}\frac{\sigma^2}{\lambda_j}\\ \mathrm{trace}(EQM(\hat{\beta}_{\mathrm{ridge}})) &= \sum_{j=1}^{p}\frac{\sigma^2\lambda_j + \lambda^2[P'\beta]_j^2}{(\lambda_j+\lambda)^2}\end{aligned}$$

Il existe toujours un λ tel que $tr[EQM(\hat{\beta}_{\mathrm{ridge}})] \leq tr[EQM(\hat{\beta}_{MC})]$

Remarques

- Il faut choisir λ
- chaque coefficient doit être pénalisé de la même manière, il faut donc que les variables associées soient du même ordre de grandeur (réduction).
- la pénalisation doit-elle intégrer l'intercept ? La constante n'est pas une 'vraie' variable et en général pas dans la contrainte donc on résoud

$$\|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_2^2$$

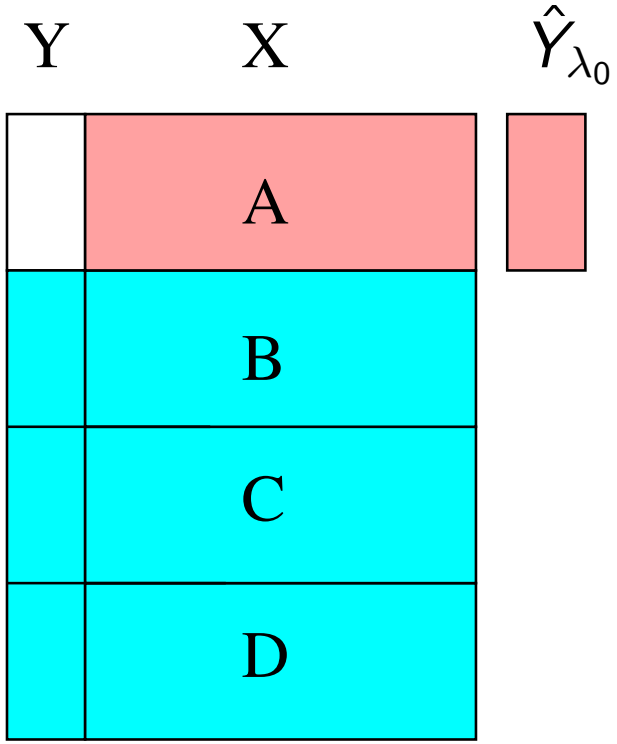
où X est la matrice des variables explicatives sans la constante (mais on conserve la même notation).

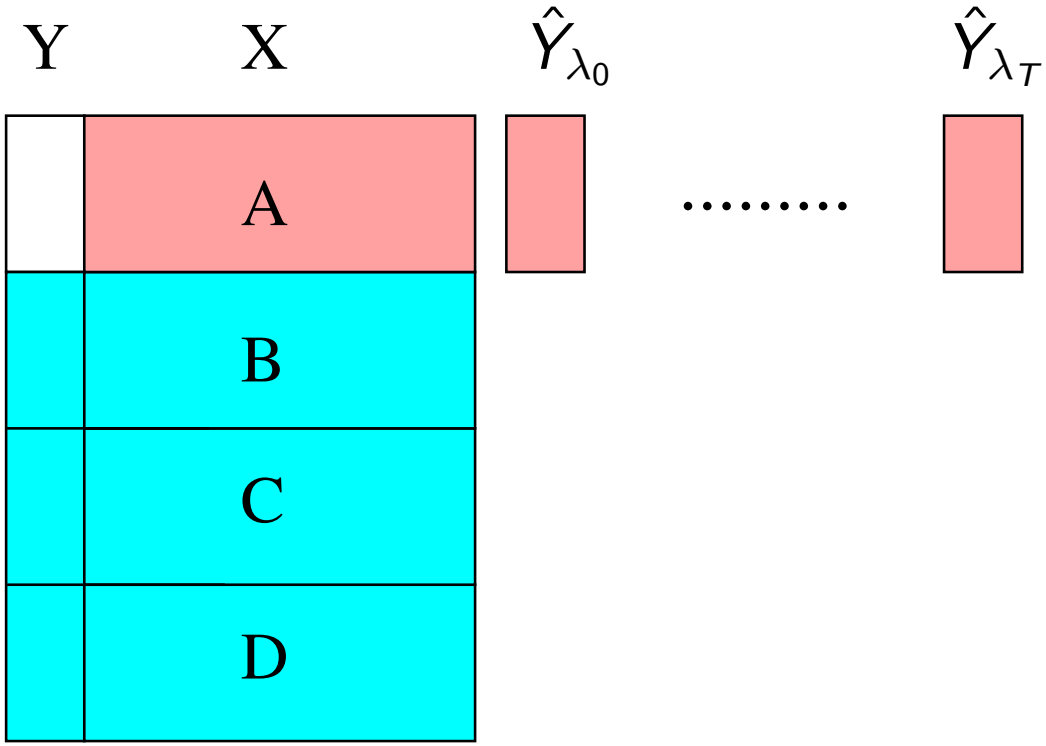
Choix de λ par VC

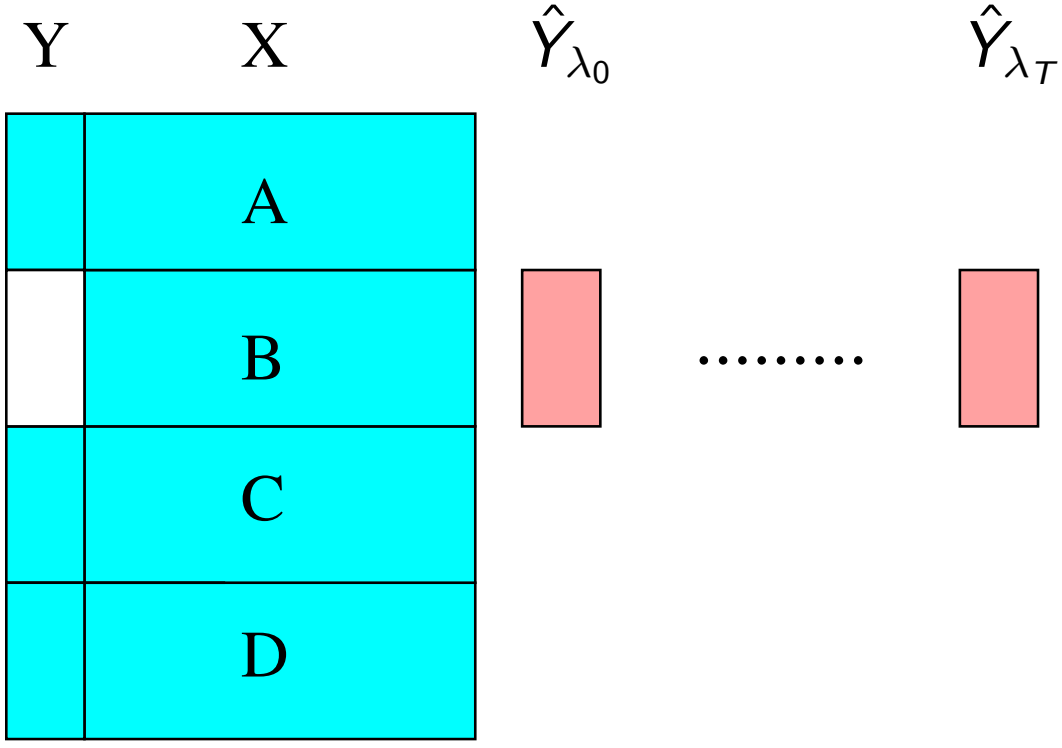
Le choix de λ est crucial. Un λ petit et nous sommes proches des MC et un λ grand et nous sommes proches de l'origine.

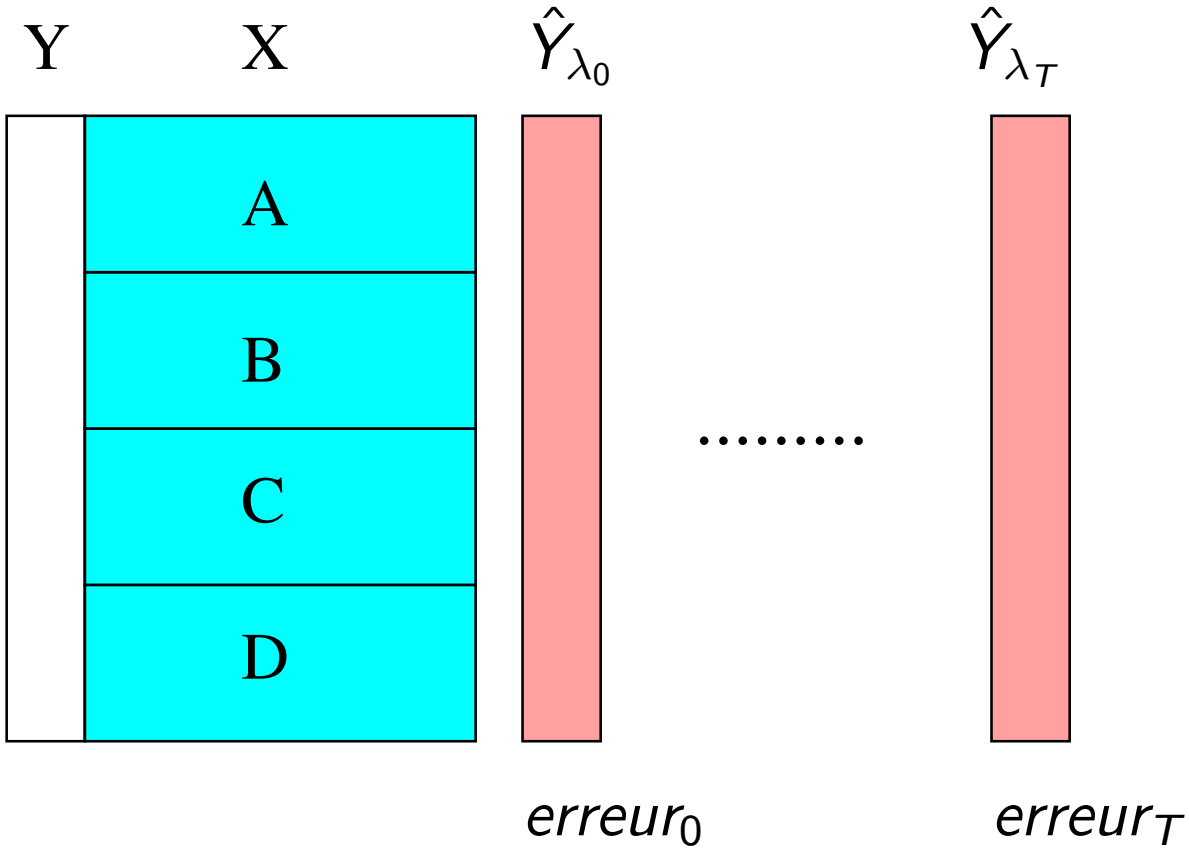
- Choix de la grille de λ : $\lambda_0, \cdots, \lambda_T$
- Choix de l'échantillon de prévision :
  - Sans : possible mais personne ne le fait
  - Apprentissage/Validation
  - Validation croisée par blocs : on découpe les données en K Blocs, on utilise $K-1$ blocs pour estimer les paramètres et on prévoit le bloc non utilisé

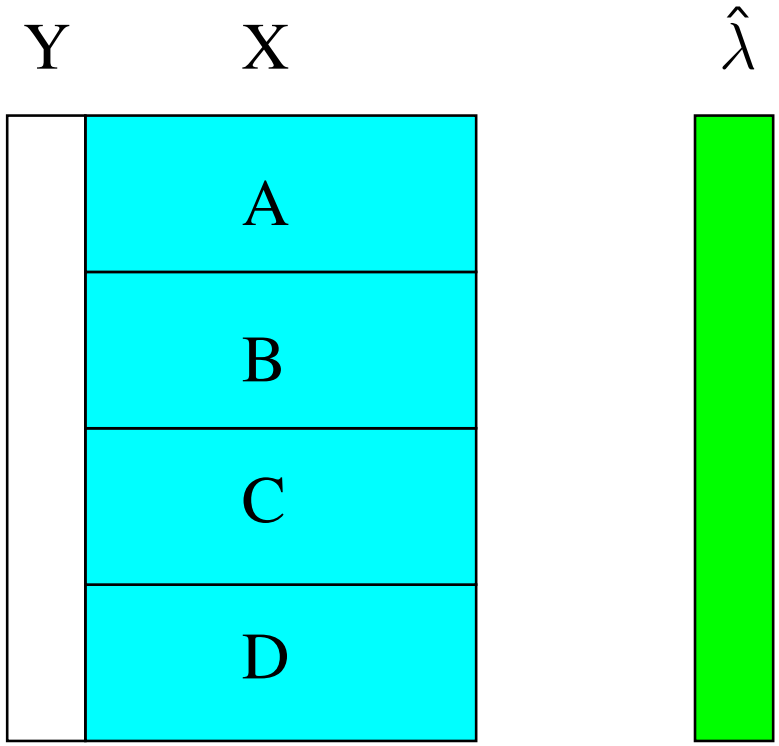
Séparation




Y
X
$\hat{Y}_{\lambda_0}$
$\hat{Y}_{\lambda_T}$
A
B
C
D
.....


Y
X
$\hat{Y}_{\lambda_0}$
$\hat{Y}_{\lambda_T}$
A
B
C
D
.....


Y
X
$\hat{Y}_{\lambda_0}$
$\hat{Y}_{\lambda_T}$
A
B
C
D
.....
$erreur_0$
$erreur_T$



On choisit $\hat{\lambda}$ et on a alors $\hat{\beta}_{\text{ridge}}(\hat{\lambda})$ avec toutes les données.

Et si on changeait de norme pour la contrainte ?

Régression Lasso

Objectif

Nous avons supposé que le modèle de régression

$$Y = X\beta + \varepsilon$$

était correct et que la matrice X était de plein rang.

On a vu la régression ridge qui minimisait à un λ donné

$$\|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_2^2$$

Et si on changeait de norme ?

Définition du Lasso

$$\min_{\beta} \|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_1$$

où X est la matrice des variables explicatives sans la constante (mais on conserve la même notation) centrée réduite ? ? ? ?

- Si $\lambda = 0$ on retrouve les estimateurs de MC.
- Si λ grand, on va forcer à avoir $\hat{\beta}$ petit afin que le terme de droite soit petit.

Conclusion : on va rétrécir les estimateurs $\hat{\beta}$ et $\hat{Y}(\lambda)$ se rapprochent de l'origine.

Régression sous contrainte

$$\min_{\beta} \|Y - \mu\mathbb{1} - X\beta\|^2 + \lambda\|\beta\|_1$$

peut aussi s'écrire sous la forme suivante

$$\tilde{\beta}(\delta) = \underset{\beta\in\mathbb{R}^p, \|\beta\|_1\leq\delta}{\operatorname{argmin}} \|Y - \mu\mathbb{1} - X\beta\|^2.$$

Pas de solution explicite algorithmes pour trouver la solution LARS (least angle regression subset)...

Illustration géométrique

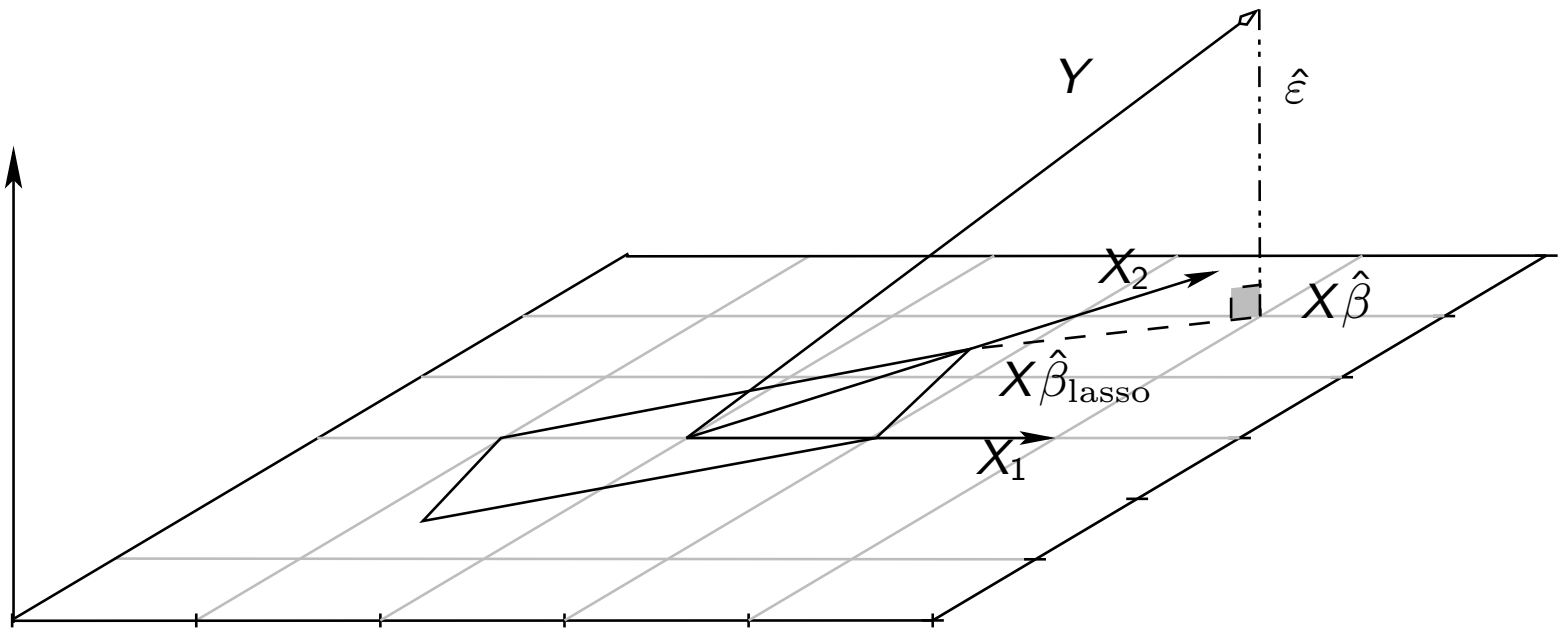
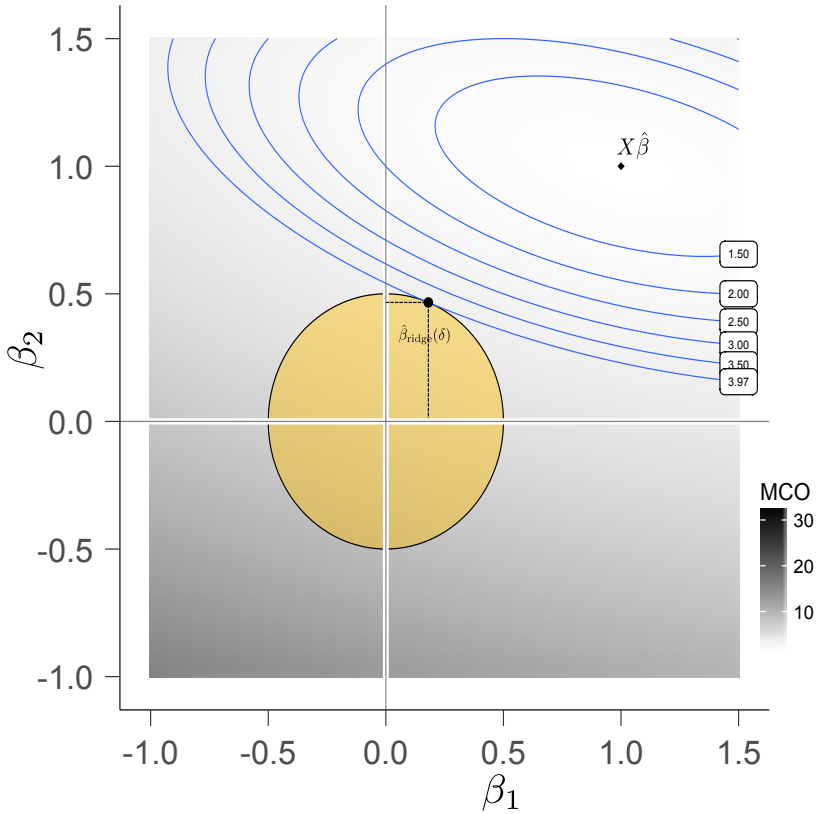
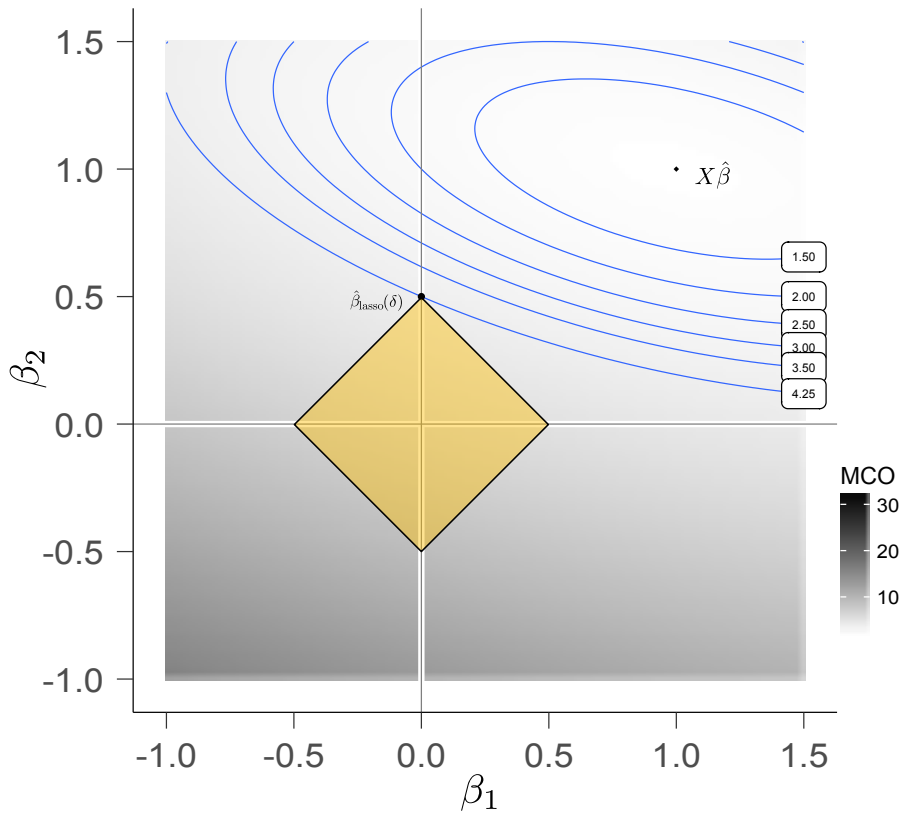


Illustration géométrique ridge/lasso





Elastic net

Un combinaison de Ridge et de Lasso

$$\hat{\beta}_{\mathrm{E}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \| Y - \mu \mathbb{1} - X\beta \|^2 + \lambda(\alpha \|\beta\|_2 + (1-\alpha)\|\beta\|_1) \right\}.$$

Pas de solution explicite algorithmes pour trouver la solution

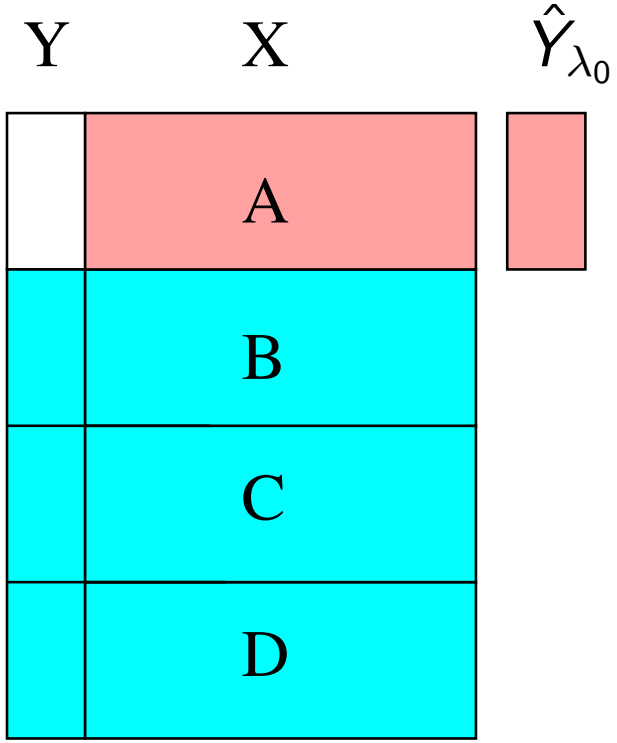
Avantage quand beaucoup de variables explicatives corrélées.

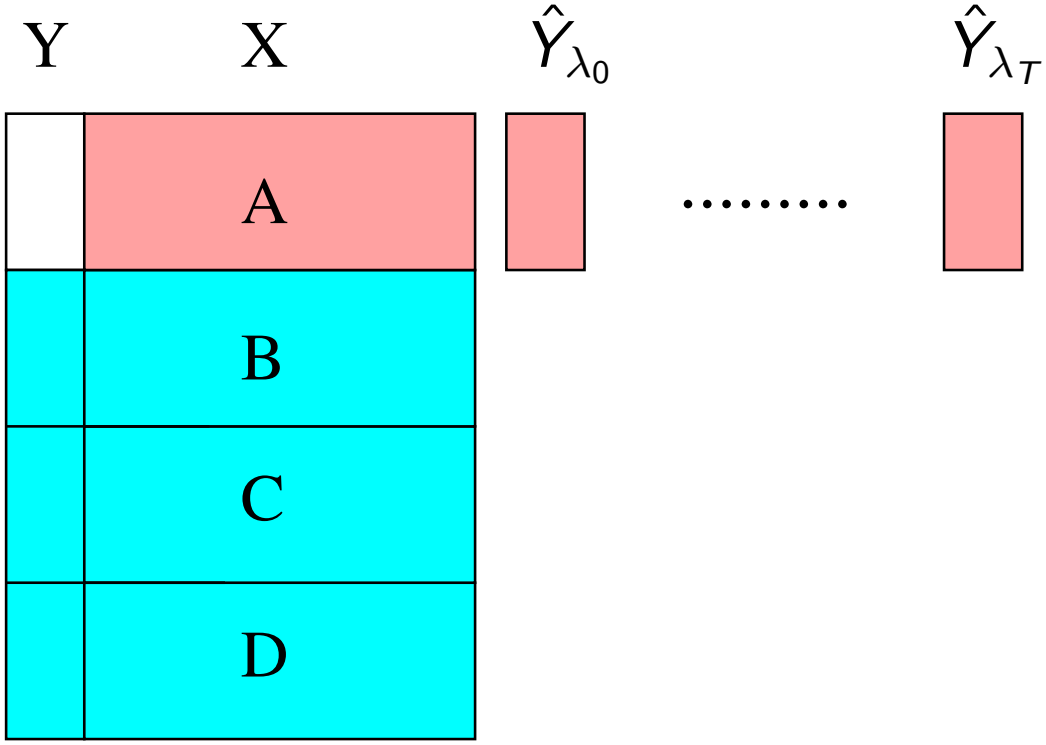
Choix de λ par VC

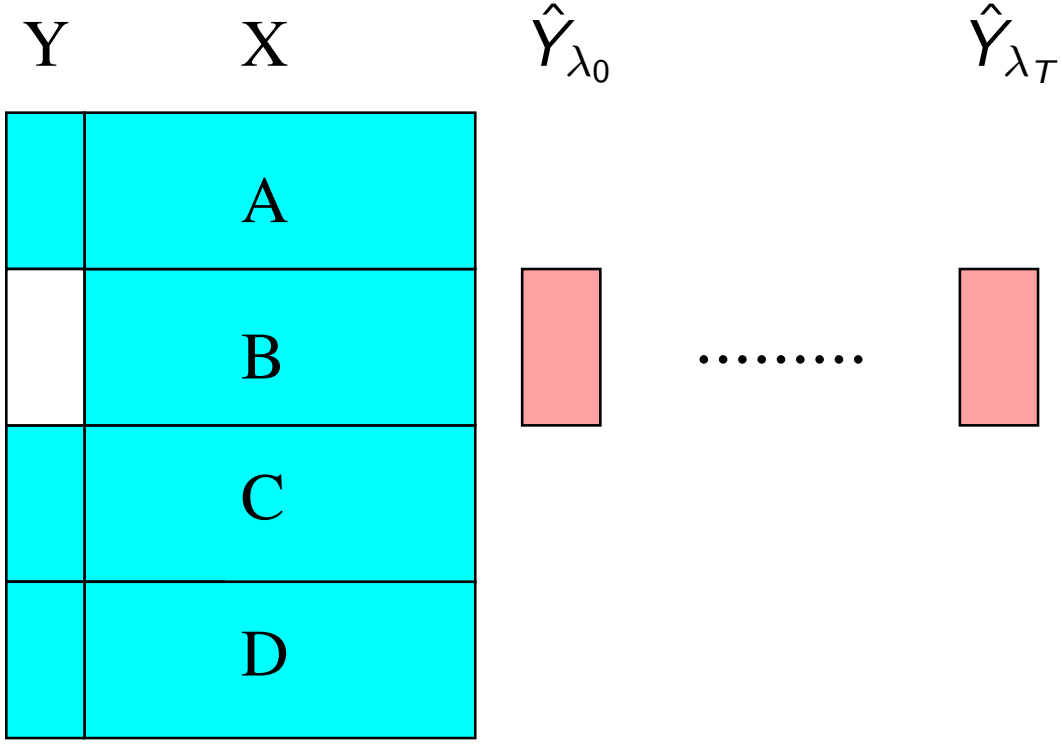
Le choix de λ est crucial. Un λ petit et nous sommes proches des MC et un λ grand et nous sommes proches de l'origine.

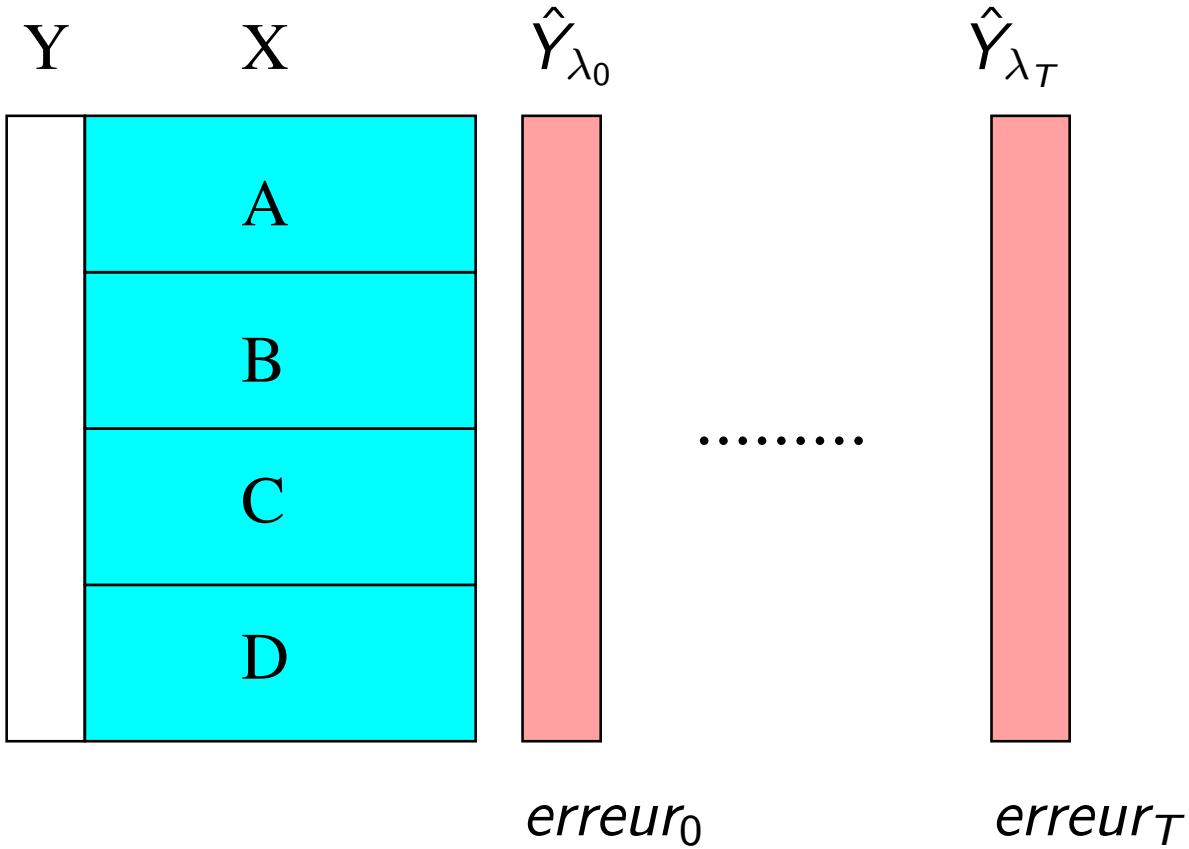
- Choix de la grille de λ : $\lambda_0, \cdots, \lambda_T$
- Choix de l'échantillon de prévision :
  - Sans : possible mais personne ne le fait
  - Apprentissage/Validation
  - Validation croisée par blocs : on découpe les données en K Blocs, on utilise $K-1$ blocs pour estimer les paramètres et on prévoit le bloc non utilisé

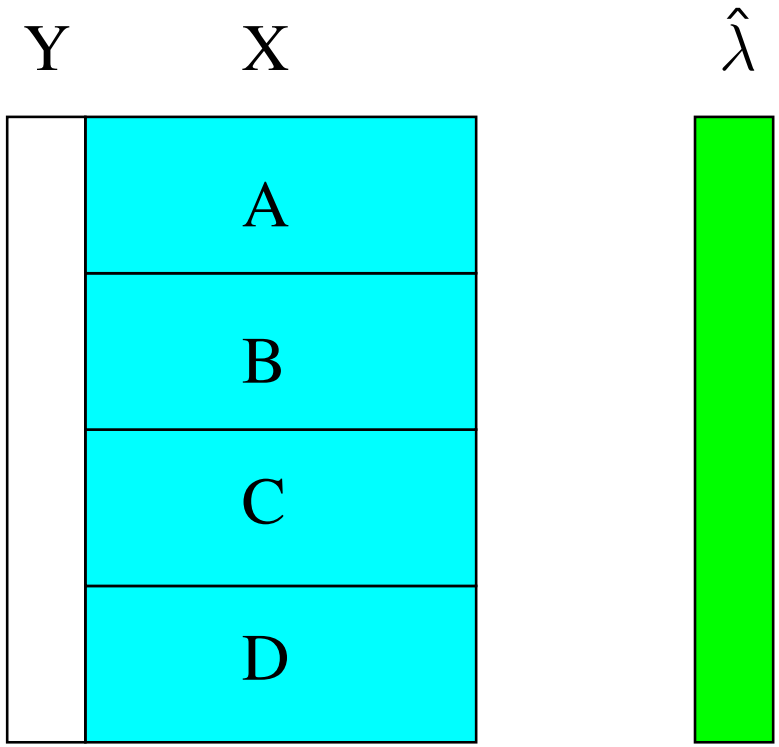
Séparation




Y
X
$\hat{Y}_{\lambda_0}$
$\hat{Y}_{\lambda_T}$
A
B
C
D
.....


Y
X
$\hat{Y}_{\lambda_0}$
$\hat{Y}_{\lambda_T}$
A
B
C
D
.....


Y
X
$\hat{Y}_{\lambda_0}$
$\hat{Y}_{\lambda_T}$
A
B
C
D
.....
$erreur_0$
$erreur_T$



On choisit $\hat{\lambda}$ et on a alors $\hat{\beta}(\hat{\lambda})$ avec toutes les données.

Comparaison de méthodes

Problème à résoudre

Nous avons présenté

- Les Moindres carrés
- Des techniques de choix de variables
- Les régressions pénalisées : ridge, lasso, elasticnet

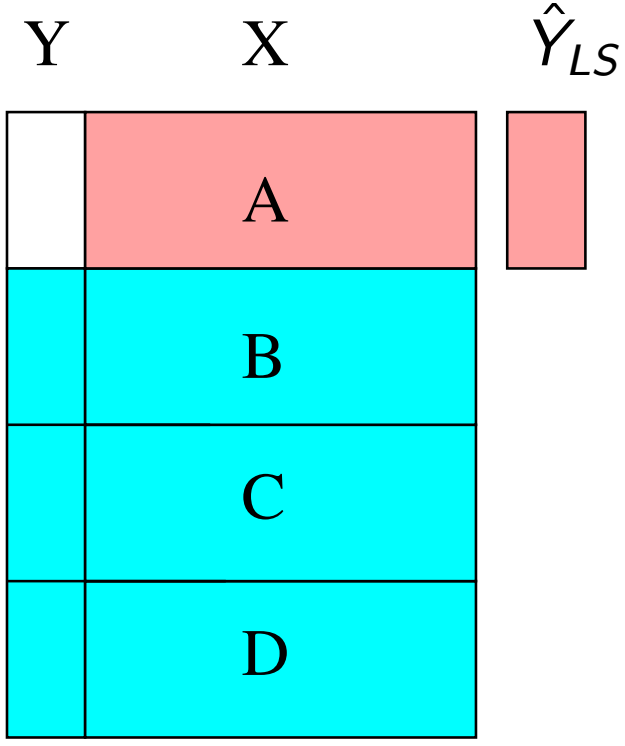
Quelle méthode choisir ?

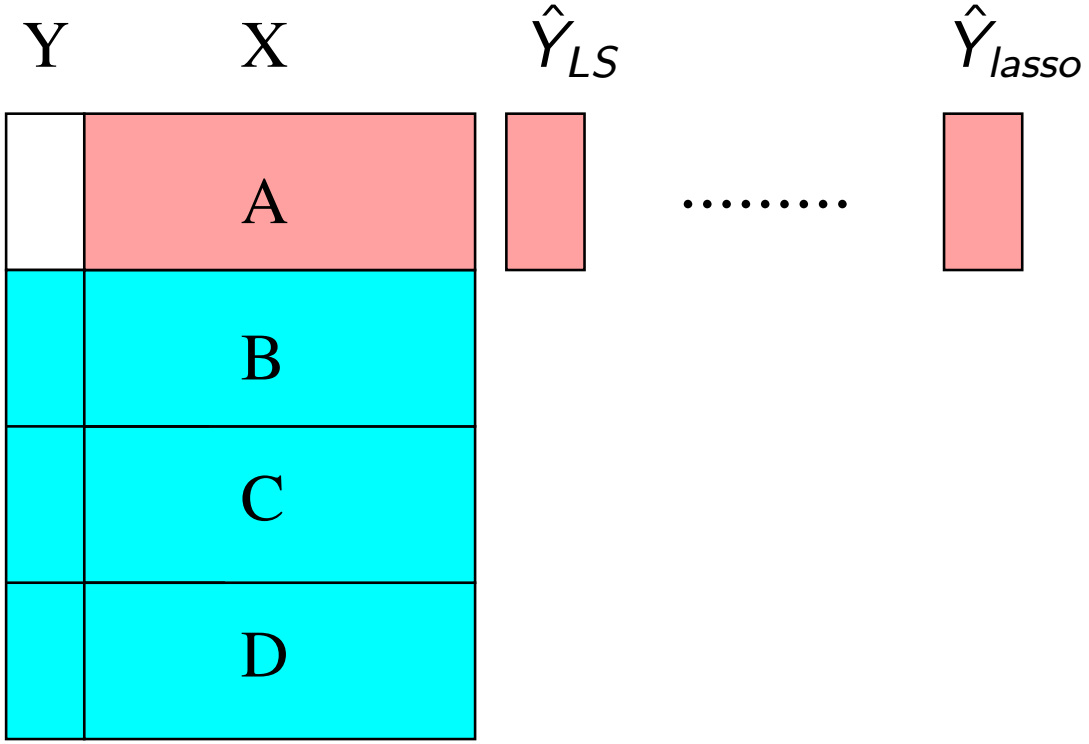
Objectif

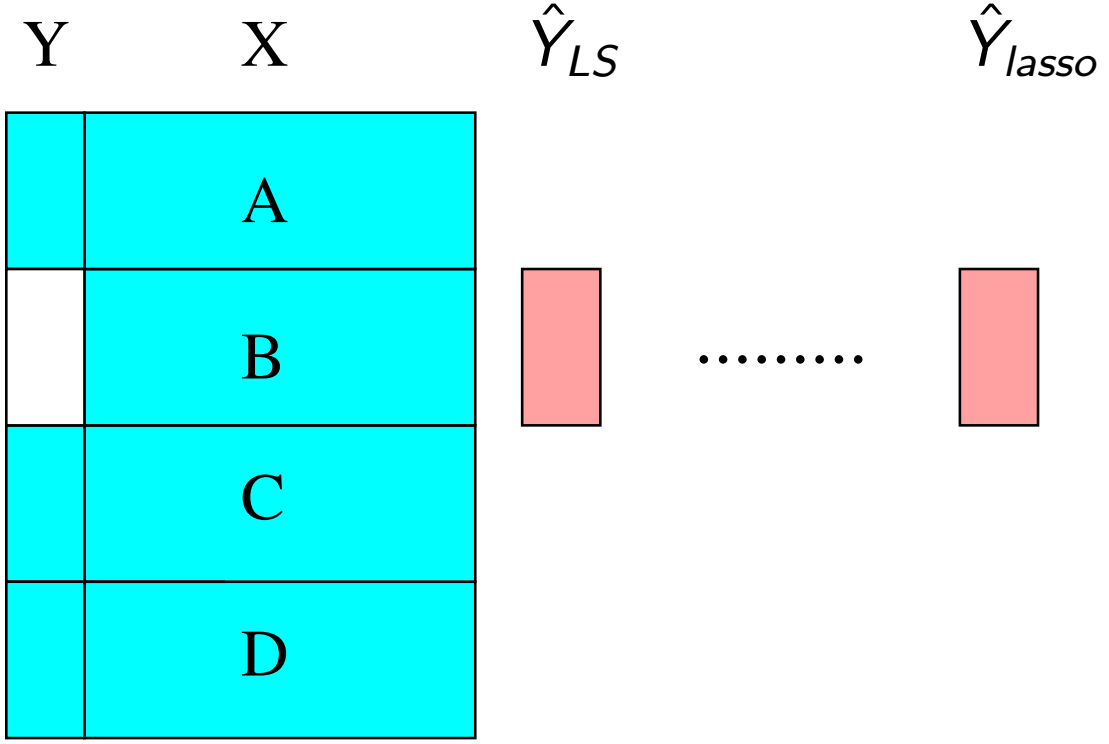
Et si on comparait sur des erreurs de prévision ?

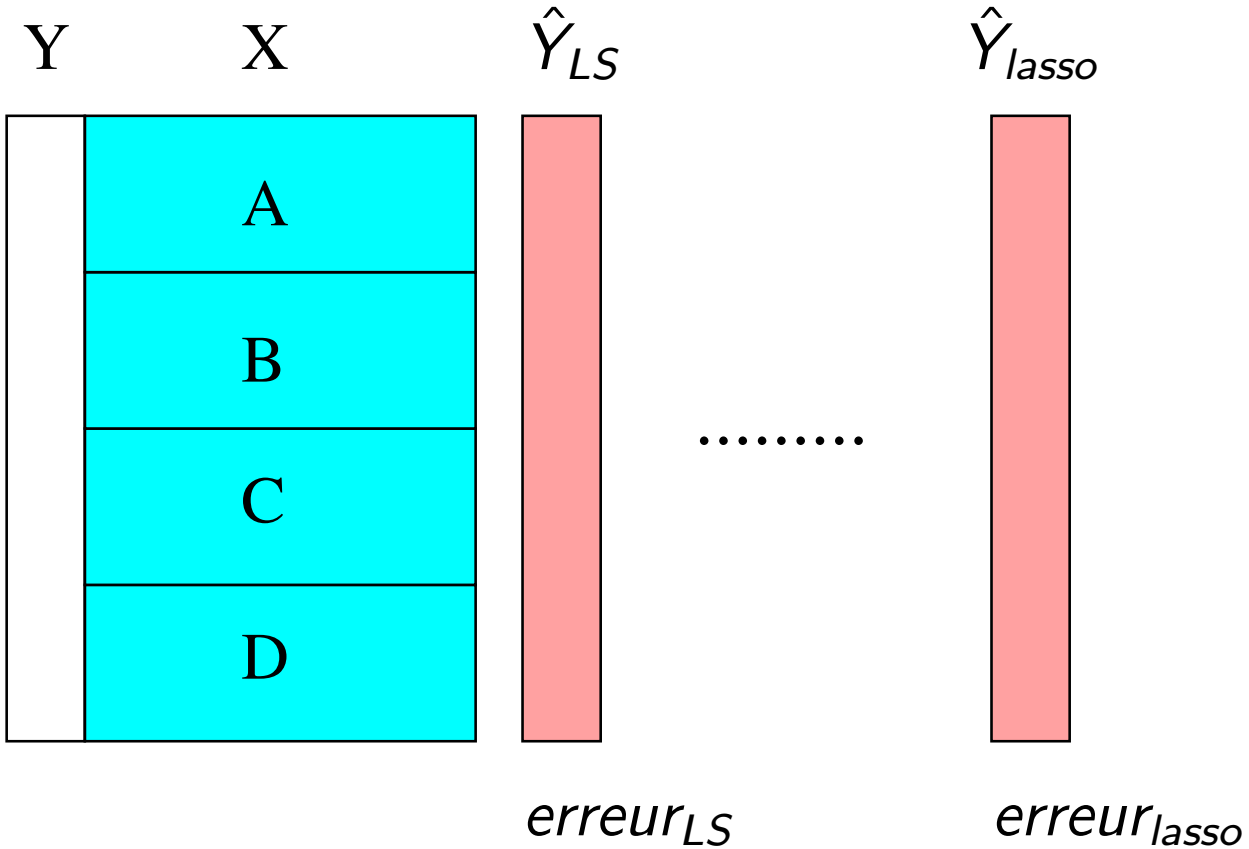
Validation croisée par blocs : on découpe les données en K Blocs, on utilise $K-1$ blocs pour estimer les paramètres et on prévoit le bloc non utilisé

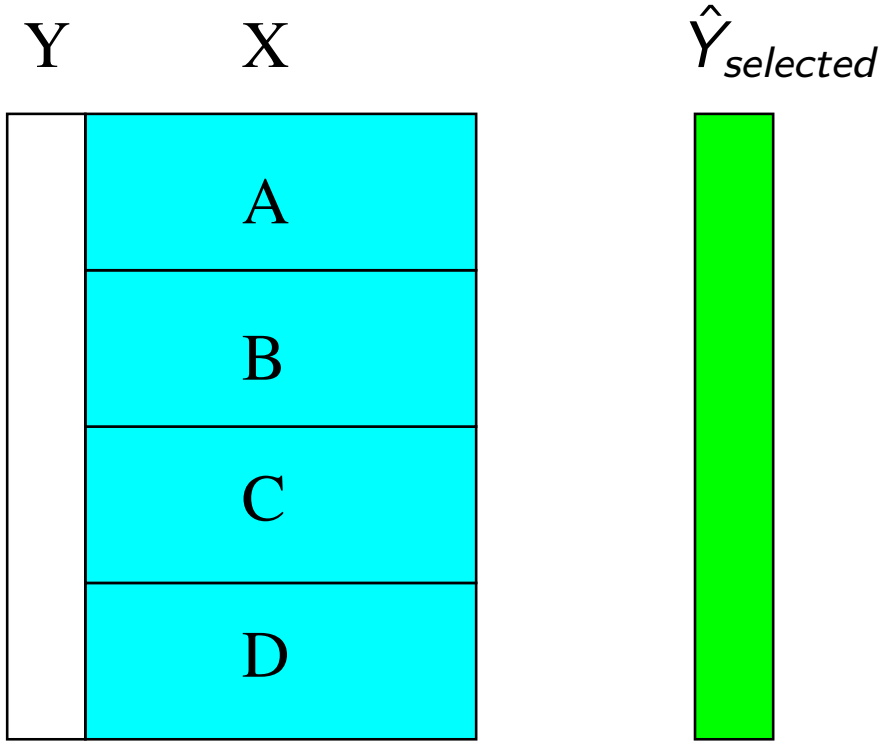
Séparation




Y
X
$\hat{Y}_{LS}$
$\hat{Y}_{lasso}$
A
B
C
D
.....


Y
X
$\hat{Y}_{LS}$
$\hat{Y}_{lasso}$
A
B
C
D


Y
X
$\hat{Y}_{LS}$
$\hat{Y}_{lasso}$
A
B
C
D
.....
$erreur_{LS}$
$erreur_{lasso}$



On choisit le meilleur algorithme en évaluant tout le monde. Puis on l'estime sur tout le monde.

Nouveaux problèmes

- **Expliquer la présence/absence d'une maladie cardio vasculaire (notée aussi CHD), par l'âge X des patients**

- Prédire l'état d'une machine outil (fonctionnement/arrêt) en fonction de son ancienneté afin de faire de la maintenance prédictive par exemple
- Analyser les espèce d'Iris : setosa, versicolor et virginica, en fonction de la longueur et largeur des pétales

Régression/Régression logistique

- Régression : expliquer une variable quantitative Y par d'autres
- Régression logistique : expliquer une variable qualitative Y par d'autres

Restriction

Seule les variables Y qualitatives binaires sont envisagées ici

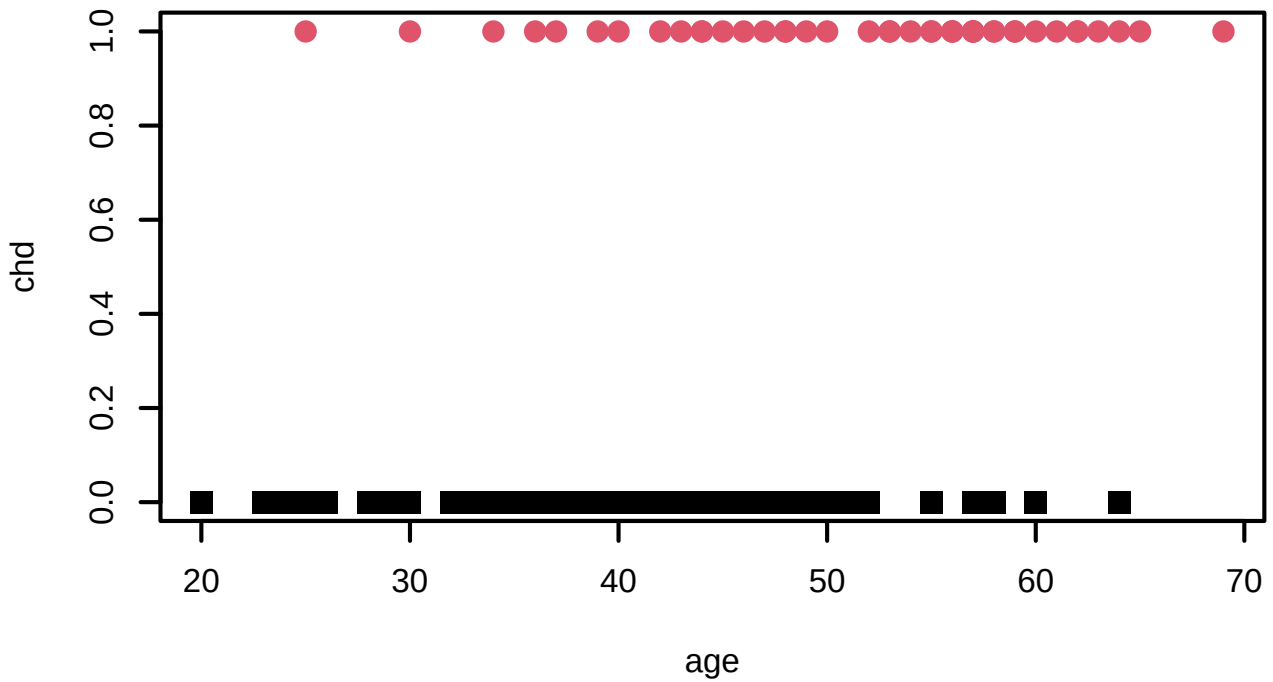
La seule différence entre ce que nous avons déjà vu est que la variable Y est qualitative et donc les outils vont changer mais la démarche va être la même.

Exemple : maladie cardio-vasculaire

- X l'âge des patients possiblement explicative.
- Y sain / malade d'une maladie cardio-vasculaire.
- $n = 100$ observations

Id	age	chd
1	20	sain
2	23	sain
3	24	sain
4	25	malade
⋮		⋮
97	64	sain
98	64	malade
99	65	malade
100	69	malade

Représentation graphique

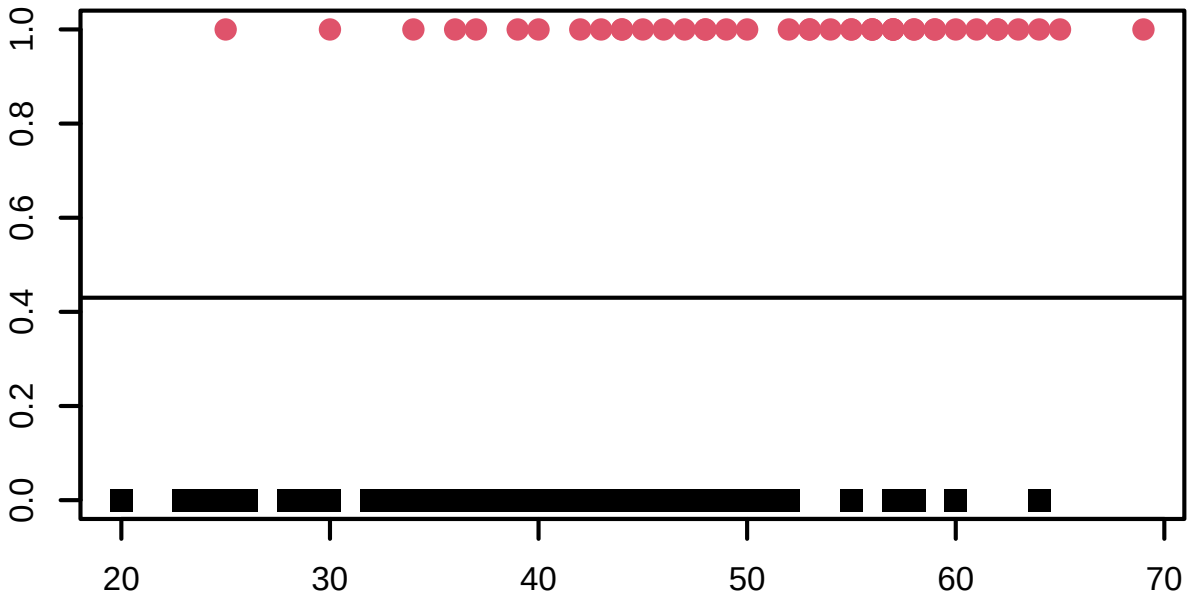


Approche naïves

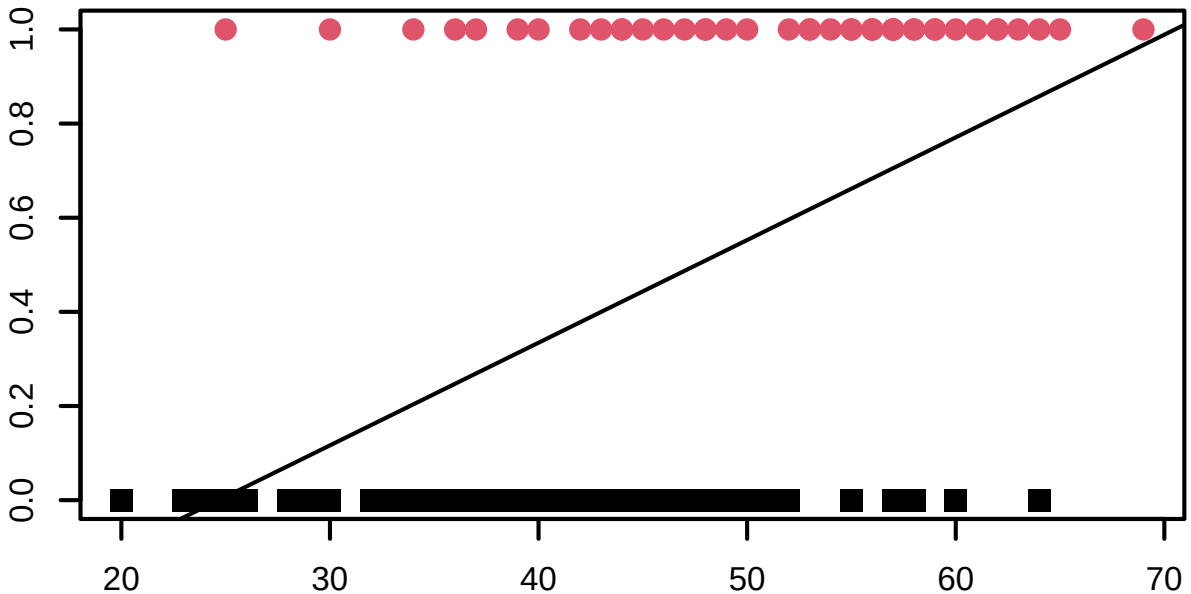
- Donner une proportion globale de malades (et l'âge ne sert pas)
- Faire une régression linéaire mais il faut coder les modalités *sain* et *malade* par exemple 0,1 ou -1,1.
- Donner une proportion de malades par classes d'âges et il faut choisir les classes
- ...

On estime une proportion, mais il faudra bien repasser aux labels initiaux sain et malade.

Pas d'effet âge



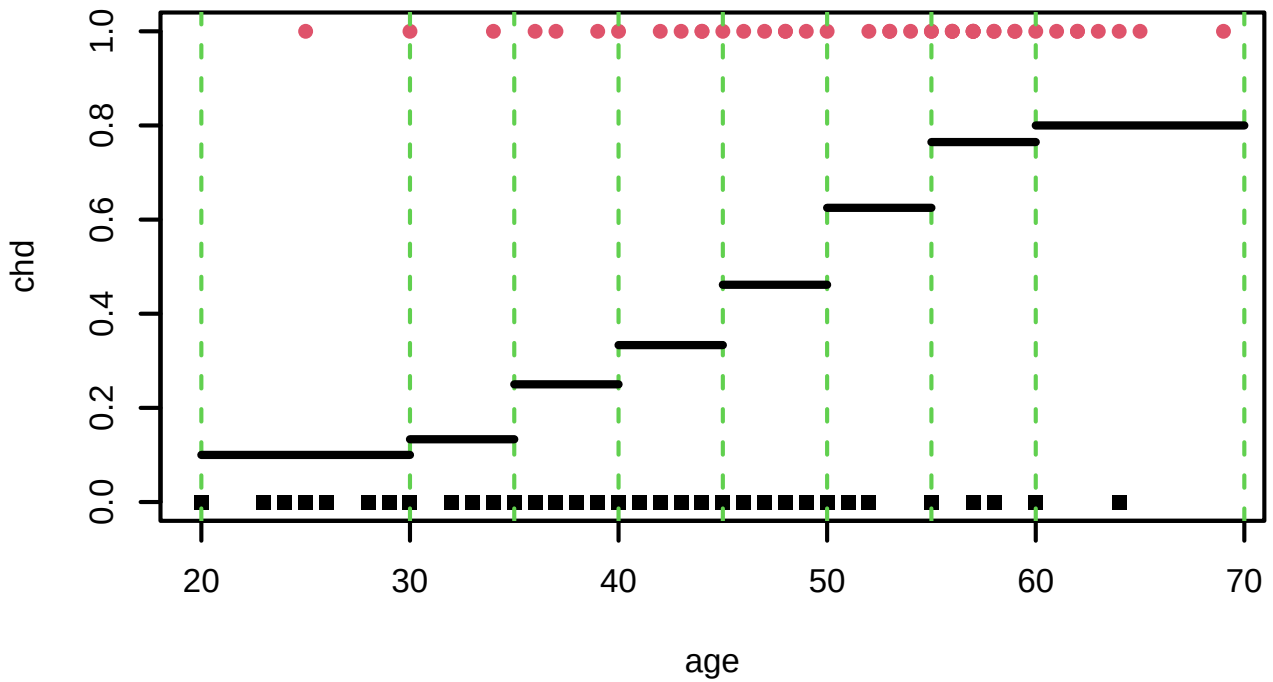
Régression linéaire



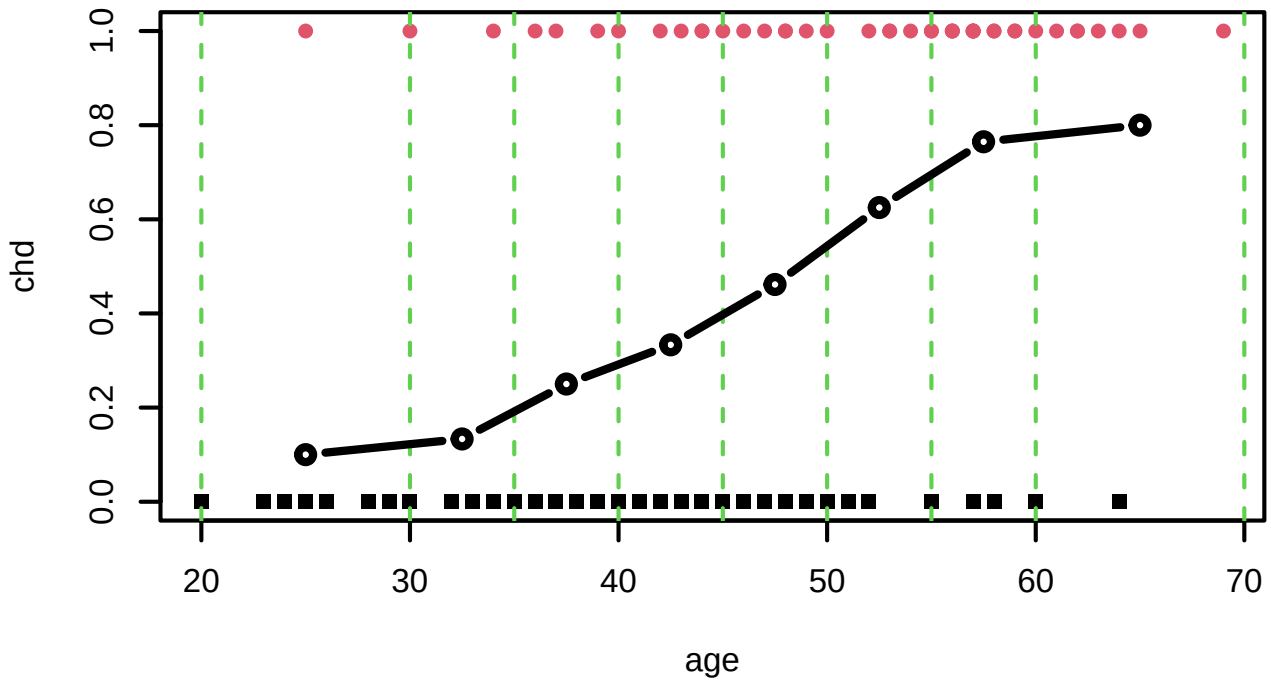
Première estimation, découpage en classes d'âge

Age	n	Absent	Présent	Moyenne
$[20, 30[$	10	9	1	.10
$[30, 35[$	15	13	2	.13
$[35, 40[$	12	9	3	.25
$[40, 45[$	15	10	5	.33
$[45, 50[$	13	7	6	.46
$[50, 55[$	8	3	5	.625
$[55, 60[$	17	4	13	.76
$[60, 70]$	10	2	8	.8

Représentation graphique



- On perd la continuité
- Comment repasse-t-on à sain et malade ?

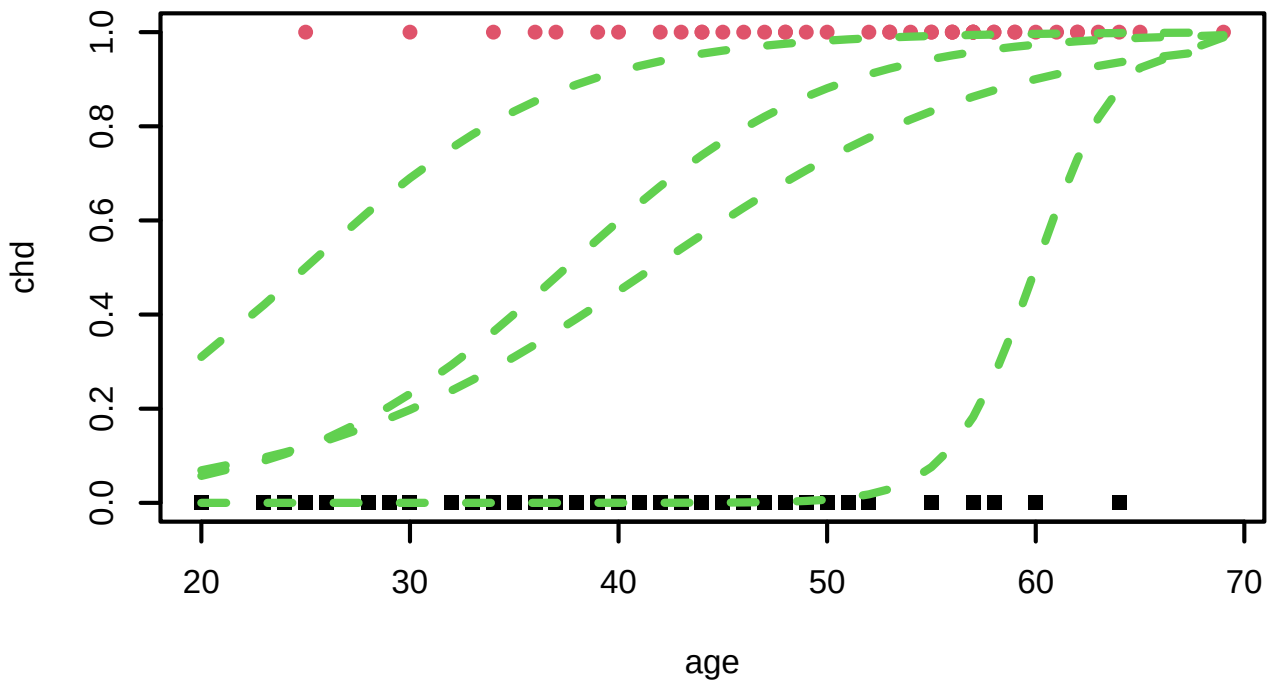

chd
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
age

Représentation graphique

- On aimerait une fonction plus lisse qui dépende de toutes les données.
- On aimerait ensuite repasser à sain et malade
  - On estime la probabilité d'être malade
  - On choisit un seuil (en général 0.5) pour passer de la probabilité à l'étiquette

Fonction souhaitée sigmoïde ?

$$f(age) = \frac{\exp(\beta_1 + \beta_2 age)}{1 + \exp(\beta_1 + \beta_2 age)}.$$



Problèmes

- Quelle fonction de coût doit-on minimiser ?
  ou
  Quelle vraisemblance doit-on maximiser ?
- Comment revenir aux labels initiaux ?

Y variable binaire

Ici la variable Y prend 2 valeurs suit donc une loi de Bernouilli avec son paramètre qui dépend (peut-être) de la variable explicative (l'âge).

$$(Y|X = x) \sim \mathcal{B}(p(x))$$

$$\mathbb{P}(Y = 1|X = x) = p(x) \quad \text{et} \quad \mathbb{P}(Y = 0|X = x) = 1 - p(x)$$

Et on suppose

$$p(x) = \frac{\exp(\beta_1 + \beta_2 \times age)}{1 + \exp(\beta_1 + \beta_2 \times age)}$$

X multivarié

Si nous avons plusieurs variables potentiellement explicatives X_1 qui peut être la constante, $X_2, \cdots, X_p$, on peut généraliser l'expression de la probabilité (en utilisant la matrice X du plan d'expérience) et obtenir pour le i^e individu/ligne :

$$\begin{aligned} p(x_i) &= \frac{\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})} \\ &= \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \end{aligned}$$

Ecriture de la vraisemblance

Exprimons la vraisemblance en fonction de β :

$$\ell(Y, \beta) = \prod_{i=1}^{n} \mathbf{P}(Y = y_i | X = x_i) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}.$$

En passant au log, on obtient

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

Ecriture de la vraisemblance

On suppose que

$$p(x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

On remplace

$$\begin{aligned} L_n(\beta) &= \sum_{i=1}^{n} y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i)) \\ &= \sum_{i=1}^{n} y_i \log(\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)})) + (1-y_i)\log(1-\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}) \\ &= \sum_{i=1}^{n} y_i x_i'\beta - \log(1+\exp(x_i'\beta)) \end{aligned}$$

Maximisation de la vraisemblance

On calcule les dérivées partielles et on les annule pour obtenir les équations normales

Malheureusement...

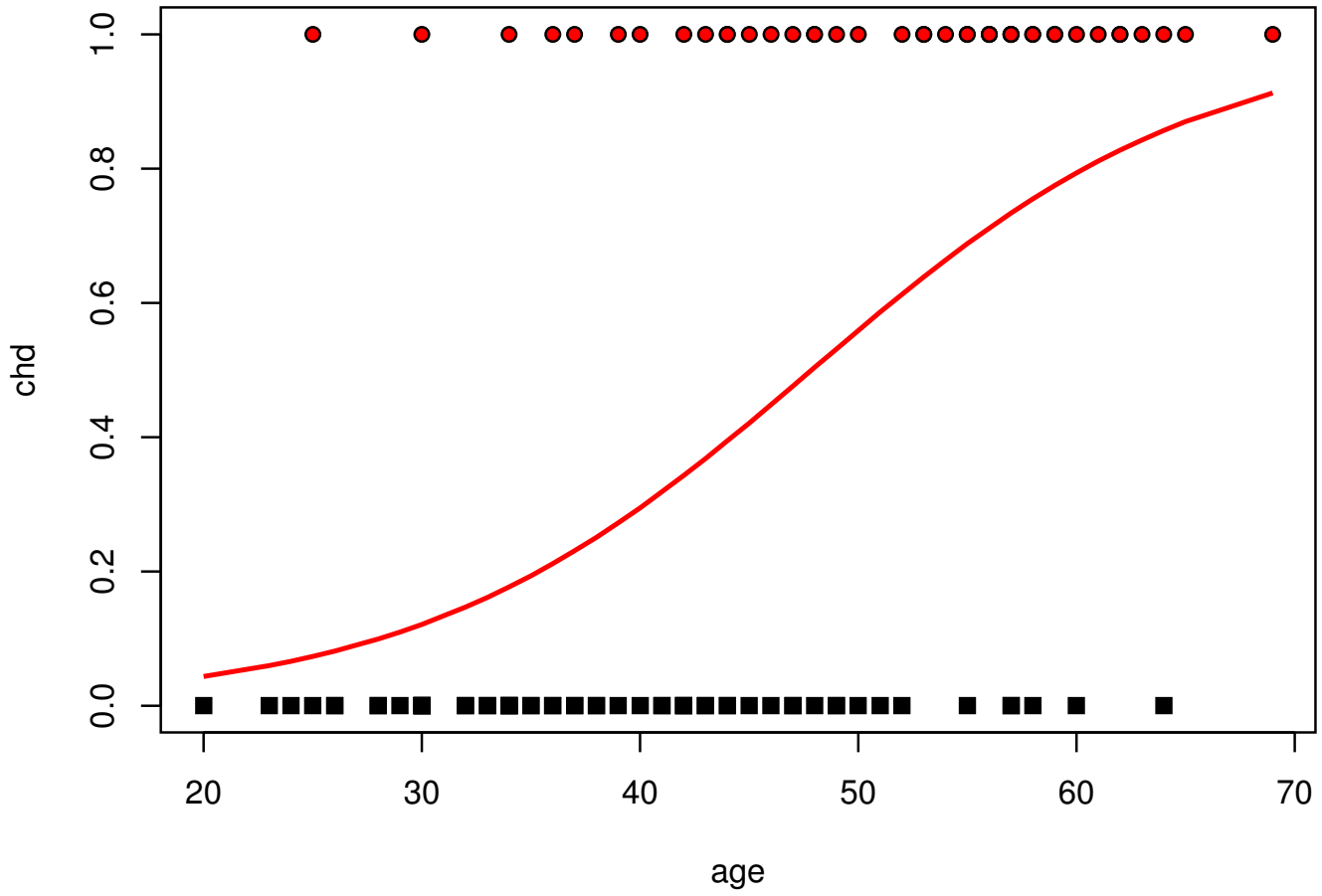
Il n'existe pas de solutions explicites pour maximiser la vraisemblance (donc pas d'écriture explicite comme pour $\hat{\beta}$).

Mais

La vraisemblance possède (généralement) un unique maximum, et il existe des algorithmes numériques itératifs permettant d'obtenir ce maximum :

- algorithme de Newton ;
- ...

Fonction estimée



$$\hat{\mathbf{P}}(Y = 1|age) = \frac{\exp(-5.30945 + 0.11092 \times age)}{1 + \exp(-5.30945 + 0.11092 \times age)}.$$

Fonction estimée

Analyser les données `artere.txt` pour retrouver les résultats du cours.

ANalyser ensuite les données de `SAheart.data`

Propriétés de $\hat{\beta}$

Expression explicites

Pas d'expression explicite de $\hat{\beta}$, $\rightarrow$ pas de calcul explicite pour son espérance et sa variance

Approximation

Théorie du max. de vraisemblance

$$\mathbb{E}(\hat{\beta}) \rightarrow_{n\rightarrow\infty} \beta$$

$$\text{Var}(\hat{\beta}) \rightarrow_{n\rightarrow\infty} \frac{\partial^2 \mathcal{L}}{\partial \beta^2} = (X' W_\beta X)^{-1}$$

avec W diagonale $W_{ii} = p_\beta(x_i)(1 - p_\beta(x_i))$

Inférence

Sous certaines hypothèses de régularité, on peut démontrer que $\hat{\beta}_n$ est asymptotiquement gaussien et cela permet donc de calculer des intervales de confiance et de faire des tests. On pourra retrouver ces résultats dans le livre de régression avec python par exemple.

Choix de variables

1. Par apprentissage/validation
2. Par validation croisée K blocs
3. Par critères de choix AIC/BIC ou par Test (ici la loi du test est approximée)

$$\begin{aligned} AIC &= -2\mathcal{L} + 2 \times \text{nb. param.} \\ BIC &= -2\mathcal{L} + ln(n) \times \text{nb. param.} \end{aligned}$$

Choix de modèles (en régression) par critère de choix

Idée naïve

1. Estimation de tous les modèles possibles $(2^p - 1)$
2. Calcul du BIC (ou AIC ou...)
3. Sélection du modèle avec meilleur critère

Problème du temps de calcul

- approche exhaustive par algo. adapté
- approche non exhaustive par forward, backward, stepwise

Algorithme Backward

- Calcul du critère pour le modèle à p variables noté M_0.
  - Calcul du critère pour les p modèles à $p-1$.
  - Choix du meilleur modèle noté M_1
- Si $M_0 < M_1$ on conserve M_0 et on arrête.
- Si $M_1 < M_0$ le modèle M_1 devient M_0 on recommence.

Estimation/prévision

- On **reprend** un individu x_i dont le couple (x_i, y_i) a servi à estimer β et on calcule $\hat{p}_i = \frac{\exp x_i' \hat{\beta}}{1+\exp x_i' \hat{\beta}}$ on obtient un $\hat{p}_i$ **estimé**.
- On **mesure un nouvel** individu x_* on calcule alors son p^* correspondant $\hat{p}^* = \frac{\exp x^{*\prime} \hat{\beta}}{1+\exp x^{*\prime} \hat{\beta}}$ on obtient un $\hat{p}^*$ **prévu**.

Problème

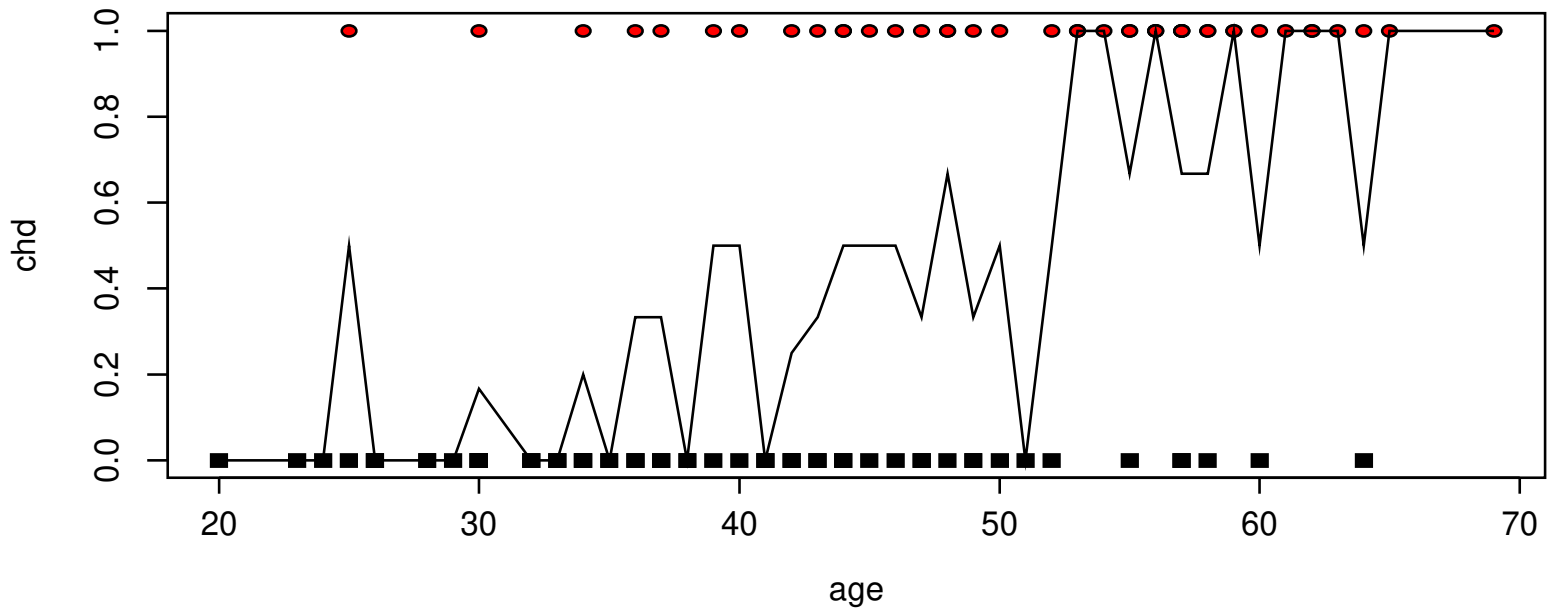
Comment calcule t'on l'erreur Y_i - " Y_i estimé" (ou prévu) ?

Modèle saturé

La vraisemblance vaut

$$\ell(Y, \beta) = \prod_{i=1}^{n} p(x_i')^{y_i}(1 - p(x_i'))^{1-y_i}$$

Le modèle saturé n'impose aucune contrainte sur p_i.
Donc si tous les x_i sont différents on a n paramètres p_i, on interpole et $\hat{p}_i = y_i$.
Si des x_i sont identiques on estime $\hat{p}_i$ par la moyenne des y_i correspondants.



On a toujours

$$\mathcal{L}(Y, \hat{\beta}) \quad \leq \quad \mathcal{L}_{\text{sat}}(Y, \hat{p})$$

Déviance

Et la déviance est définie comme

$$\mathcal{D} = -2(\mathcal{L}(Y, \hat{\beta}) - \mathcal{L}_{\text{sat}}(Y, \hat{p})) \geq 0,$$

La déviance est toujours positive. C'est un indicateur qui mesure la qualité d'ajustement du modèle : plus la déviance est faible, mieux le modèle ajuste les données.

Quel résidus ?

1. Résidus de Pearson

$$\hat{\varepsilon}_{Pi} = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

2. Résidus de déviance

$$\hat{\varepsilon}_i = \text{signe}(y_i - \hat{p}_i)\sqrt{2(\mathcal{L}_{\text{sat}}(y_i) - \mathcal{L}(y_i, \hat{\beta}))}$$

Prévision d'un label

Seuil

Afin d'obtenir un label $\hat{Y}^*$ à partir d'une probabilité $\hat{P}(Y = 1|X = x^*)$ on choisit un seuil s :

- $\hat{\mathbb{P}}(Y = 1|X = x^*) \leq s \Rightarrow \hat{Y}^* = 0$
- $\hat{\mathbb{P}}(Y = 1|X = x^*) > s \Rightarrow \hat{Y}^* = 1$

Une fois le seuil s choisi, on peut calculer soit en estimation soit en prévision

Règle de Bayes

On compare

$$\hat{p}(x_i) = \hat{\mathbb{P}}(Y = 1|X = x_i) \text{ à } \hat{\mathbb{P}}(Y = 0|X = x_i) = 1 - p(x_i).$$

On choisit la plus grande valeur/le plus probable : $s = 0.5$

Mais est-ce que ce choix est réaliste ?

Matrice de confusion estimée ou prédite

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	VN	FP
$Y=1$	FN	VP

	$\hat{Y}^*=0$	$\hat{Y}^*=1$
$Y^*=0$	VN	FP
$Y^*=1$	FN	VP

- les VP vrais positifs représentent les malades admettant un test positif
- les FP faux positifs représentent les non-malades admettant un test positif
- les FN faux négatifs représentent les malades admettant un test négatif
- les VN vrais négatifs représentent les non-malades admettant un test négatif

Matrice de confusion estimée ou prédite

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	VN	FP
$Y=1$	FN	VP

	$\hat{Y}^*=0$	$\hat{Y}^*=1$
$Y^*=0$	VN	FP
$Y^*=1$	FN	VP

On définit :

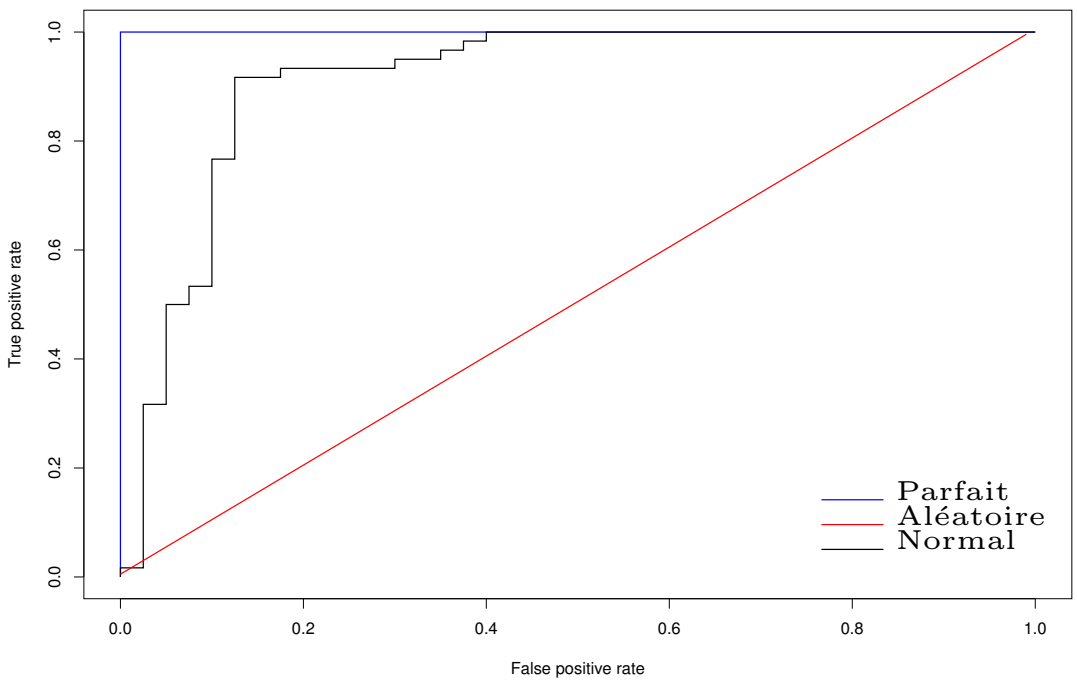
- Accuracy : $(VN+VP)/n$
- Spécificité : $sp(s)=\mathbb{P}(S(X)<s|Y=0)$ probabilité que le test soit négatif quand la maladie est non présente : $VN/(VN+FP)$.
- Sensibilité : $se(s)=\mathbb{P}(S(X)\geq s|Y=1)$ probabilité que le test soit positif quand la maladie est présente $VP/(VP+FN)$.

Remarque : la sensibilité sans la spécificité cela ne sert à rien : imaginez un test qui est toujours positif.

Courbe ROC

C'est une courbe paramétrée par le seuil :

$$\begin{cases} x(s) = 1 - sp(s) = \mathbb{P}(S(X) > s | Y = 0) \\ y(s) = se(s) = \mathbb{P}(S(X) \geq s | Y = 1) \end{cases}$$



Vraisemblance pénalisée

$$\begin{aligned}\mathcal{L}(\beta, \lambda) &= \sum_{i=1}^{n} \{y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))\} - \lambda \|\beta\| \\ &= \sum_{i=1}^{n} \{y_i(\mu + x_i'\beta) - \log(1 + \exp(\mu + x_i'\beta))\} - \lambda \|\beta\|\end{aligned}$$

Et donc l'estimation de μ varie avec λ.

Comparaison de méthodes

Problème à résoudre

Nous avons présenté

- La régression logistique
- Des techniques de choix de variables
- Les régressions pénalisées : ridge, lasso, elasticnet

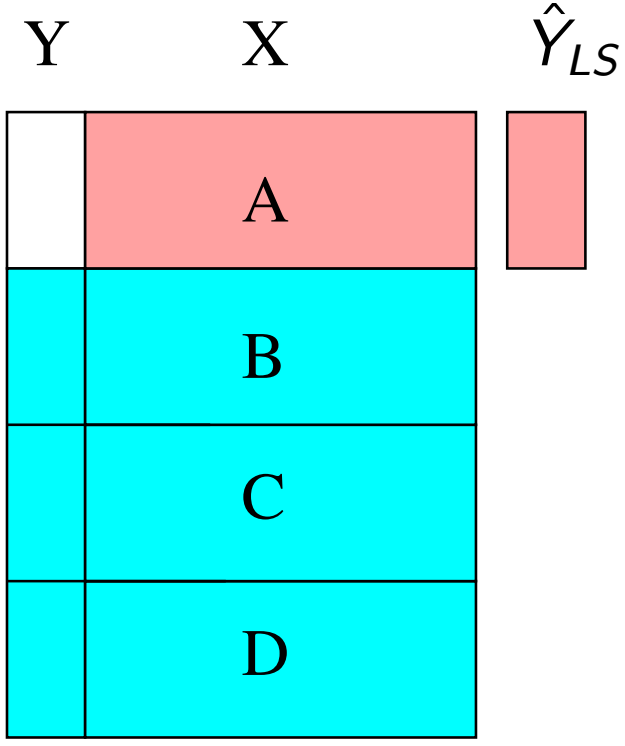
Quelle méthode choisir ?

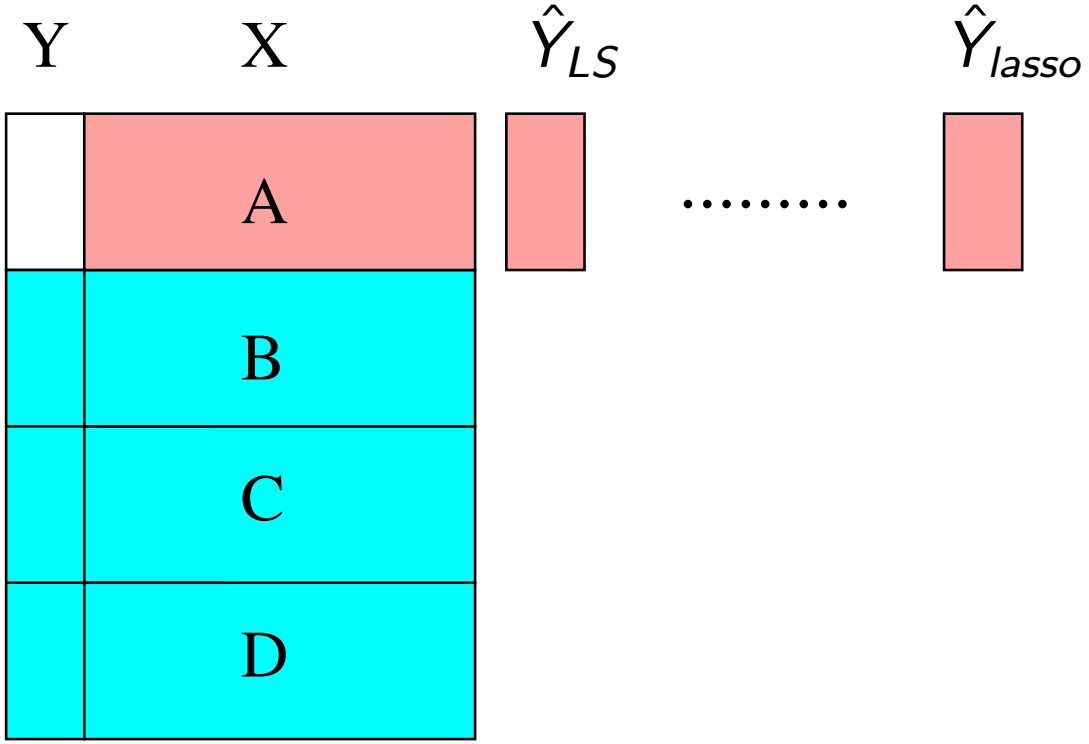
Objectif

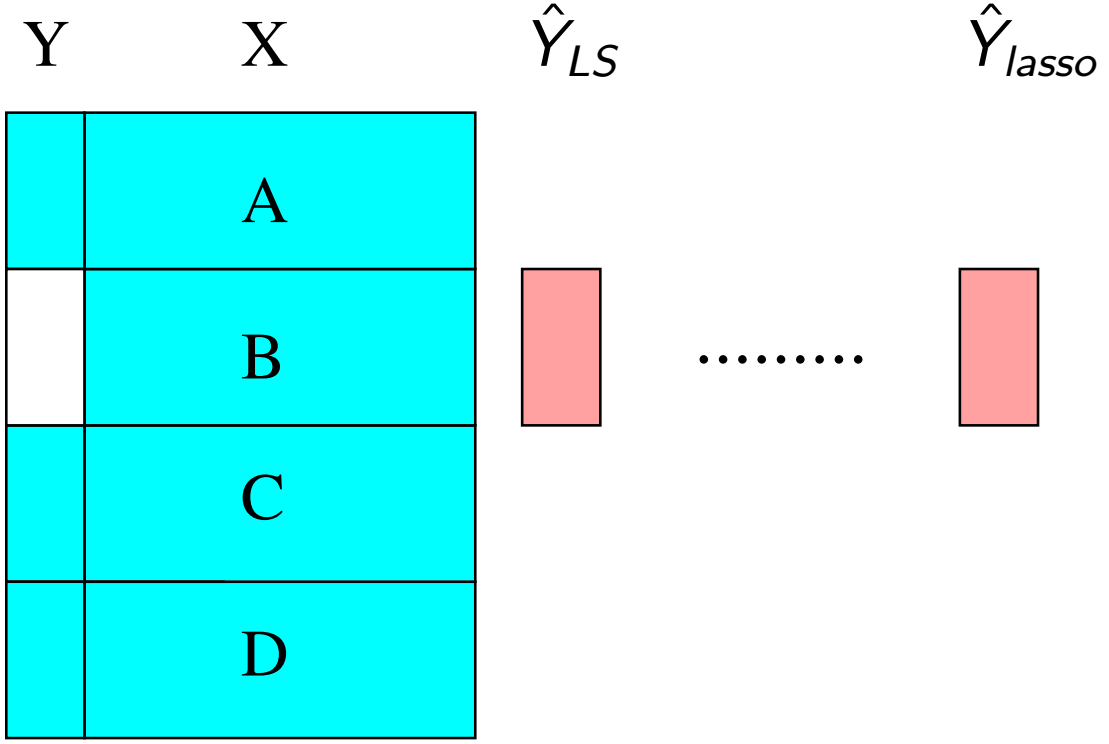
Et si on comparait sur des erreurs de prévision ?

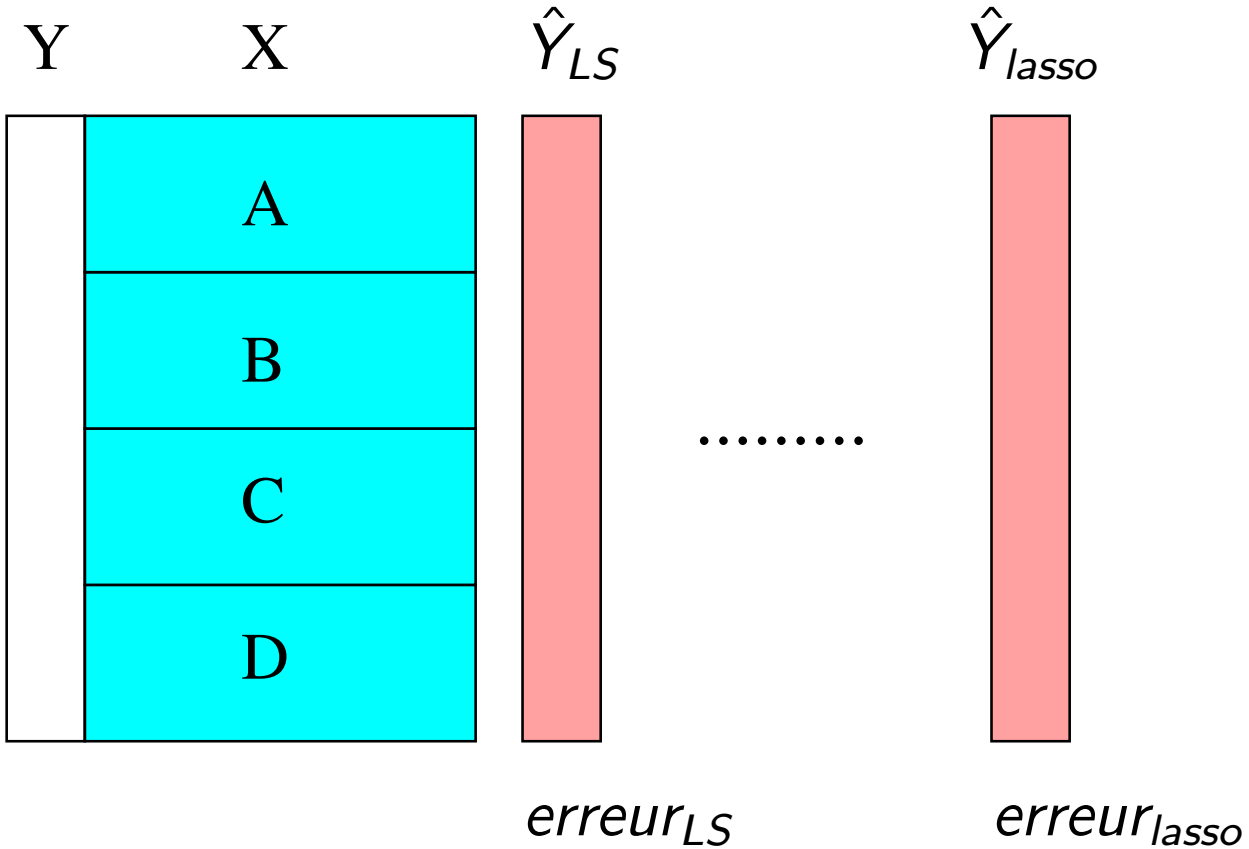
Validation croisée par blocs : on découpe les données en K Blocs, on utilise $K-1$ blocs pour estimer les paramètres et on prévoit le bloc non utilisé.
C'est exactement comme en régression sauf que nous n'avons pas de $\hat{Y}$ sauf si on fixe le seuil.

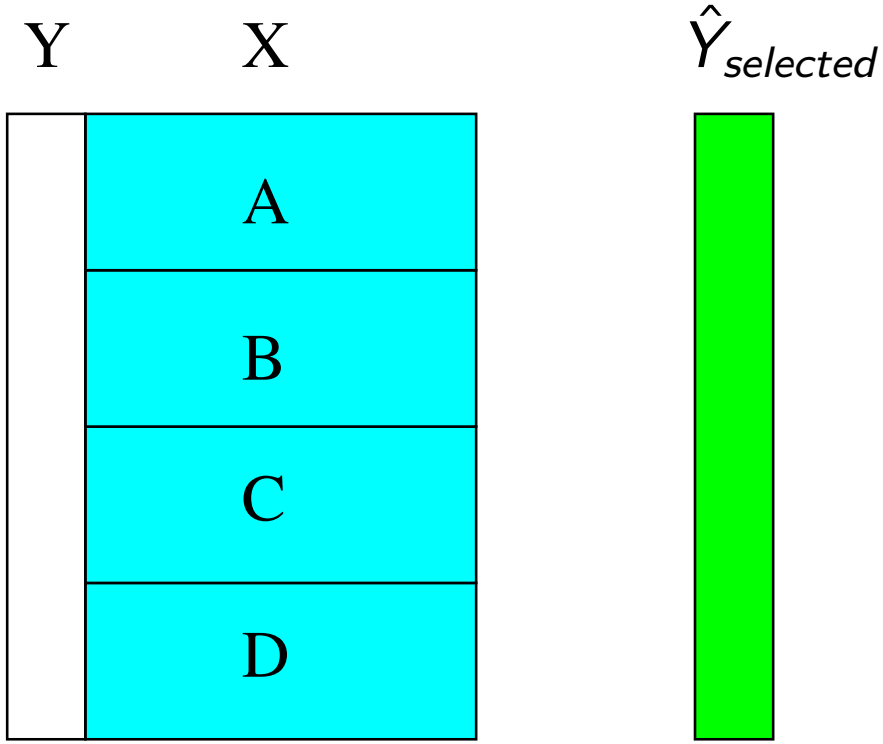
Séparation




Y
X
$\hat{Y}_{LS}$
$\hat{Y}_{lasso}$
A
B
C
D
.....


Y
X
$\hat{Y}_{LS}$
$\hat{Y}_{lasso}$
A
B
C
D


Y
X
$\hat{Y}_{LS}$
$\hat{Y}_{lasso}$
A
B
C
D
.....
$erreur_{LS}$
$erreur_{lasso}$



On choisit le meilleur algorithme en évaluant tout le monde. Puis on l'estime sur tout le monde.