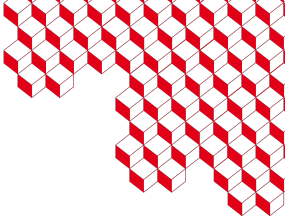


Régression linéaire et apprentissage statistique

M2 Radiophysique médicale, INSTN, 2023

Clément GAUCHY (clement.gauchy@cea.fr) Blog: clgch.github.io
CEA SACLAY

Sommaire

Introduction aux modèles linéaires

- A partir d'un ensemble d'exemples $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, on se propose d'approcher la relation entre $\mathbf{x}$ et y par un modèle (apprentissage statistique, Machine Learning)

$$y(\mathbf{x}) = \varphi(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$$

- Le modèle φ appartient à une famille de fonctions paramétriques de paramètres $\boldsymbol{\beta}$ (polynômes, fonctions de bases,...)
- La variable ε représente le bruit et/ou l'erreur par rapport au *vrai* modèle (si il existe) ayant généré les données
- Exemple des moindres carrés pour construire une meilleure approximation

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y(\mathbf{x}_i) - \varphi(\mathbf{x}_i, \boldsymbol{\beta}))^2$$

- Certaines hypothèses probabilistes sur l'erreur ε vont nous permettre d'analyser la qualité de l'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$

Modèle linéaire

- Cas où le modèle hypothèse φ est linéaire en fonction des coefficients $\boldsymbol{\beta}$:

$$y(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \underbrace{x_j}_{\text{régresseurs}} \beta_j + \varepsilon = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$$

- Exemple du modèle pour décrire la perte de charge en fonction du nombre de Reynolds (mécanique des fluides) et sa transformation en un modèle avec hypothèse linéaire

$$\begin{aligned} \Delta P &= a \times Re^{-b} \\ \ln(\Delta P) &= \ln(a) - b \times \ln(Re) \end{aligned}$$

- Coefficients et régresseurs

$$\begin{aligned} x &= \ln(Re) \\ \beta_0 &= \ln(a) \\ \beta_1 &= -b; \qquad x_1 = x \end{aligned}$$

Sommaire

Solution par moindres carrés

- Représentation matricielle très pratique

$$\mathbf{y}_{n\times 1} = \boldsymbol{X}_{n\times(p+1)} \times \boldsymbol{\beta}_{p+1\times 1} + \varepsilon_{n\times 1}$$

avec

$$\mathbf{y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} ; \quad \boldsymbol{X}_{n\times(p+1)} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p \end{pmatrix} ; \quad \varepsilon_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Solution des *moindres carrés* en supposant la matrice $\boldsymbol{X}$ de rang plein p

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \Rightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$$

- Prédicteur: $\hat{\mathbf{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$, la matrice $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ ($\boldsymbol{H}$ pour *hat matrix*) est une matrice de projection orthogonale (exercice: le vérifier).

Postulats de la régression linéaire

- P1: les erreurs sont centrées
$$\mathbb{E}[\varepsilon] = 0$$
- P2: la variance des erreurs est constante (homoscédasticité)
$$\mathrm{Var}(\varepsilon) = \sigma^2$$
- P3: les erreurs sont indépendantes
- P4: les erreurs sont Gaussiennes
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Modèle linéaire avec postulats P1 à P3 vrais

↪ L'estimation est sans biais:

$$\mathbb{E}[\hat{\beta}] = \beta \texttt{ (petit exercice !)}$$

Modèle linéaire avec postulats P1 à P3 vrais

↪ L'estimation est sans biais:

$$\mathbb{E}[\hat{\beta}] = \beta \texttt{ (petit exercice !)}$$

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}] \\
&= \mathbb{E}_{\varepsilon}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\beta + \varepsilon)] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta \\
&= \beta
\end{aligned}$$

Modèle linéaire avec postulats P1 à P3 vrais

↪ L'estimation est sans biais:

$$\mathbb{E}[\hat{\beta}] = \beta \texttt{ (petit exercice !)}$$

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}] \\
&= \mathbb{E}_{\varepsilon}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\beta + \varepsilon)] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta \\
&= \beta
\end{aligned}$$

↪ La variance de l'estimateur s'écrit

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Exercice: Vérifier que $\boldsymbol{X}^T\boldsymbol{X}$ est bien inversible

Modèle linéaire avec postulats P1 à P3 vrais

↪ L'estimation est sans biais:

$$\mathbb{E}[\hat{\beta}] = \beta \texttt{ (petit exercice !)}$$

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}] \\
&= \mathbb{E}_{\varepsilon}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\beta + \varepsilon)] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta \\
&= \beta
\end{aligned}$$

↪ La variance de l'estimateur s'écrit

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Exercice: Vérifier que $\boldsymbol{X}^T\boldsymbol{X}$ est bien inversible

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}((\boldsymbol{X}^T\boldsymbol{X}^{-1}\boldsymbol{X}^T\mathbf{y})) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Var}(\boldsymbol{X}\beta + \varepsilon)((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Var}(\varepsilon)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

... et en rajoutant le postulat P4

$\hookrightarrow$ L'estimateur des paramètres suit une loi Gaussienne

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

La matrice $\boldsymbol{X}^T\boldsymbol{X}/\sigma^2$ est appellé matrice d'information (lien direct avec la matrice d'information de Fisher)

$\hookrightarrow$ La norme des résidus suit une loi du χ^2

$$\frac{\|\mathbf{y} - \boldsymbol{X}\hat{\beta}\|^2}{\sigma^2} = \frac{\|\hat{\varepsilon}_{n\times 1}\|^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

les $\hat{\varepsilon}_{n\times 1}$ sont appelés les résidus. Pour les curieux qui veulent la preuve: voir le théorème de Cochran.

Sans P4, ces résultats ne sont vrais qu'asymptotiquement ! (Pour $n \to +\infty$)

Régression linéaire en pratique

Diagnostic

- Vérification des hypothèses: linéarité, normalité, données dites aberrantes

Transformation des données

- Transformation de la réponse y
- Transformation des variables

Sélection de variables

- Régression *stepwise*
- Critères AIC, BIC, Cp de Mallows,...

Vérification des résultats

Outil essentiel: les résidus

$$\hat{\varepsilon} = \mathbf{y} - \boldsymbol{X}\hat{\beta}$$

Vérifications graphique:

- Pour P1 et P2, valeur des résidus contre valeur prédite
- Pour P3, résidus contre temps/ordre des données

P1 & P2: adéquation et homoscédasticité

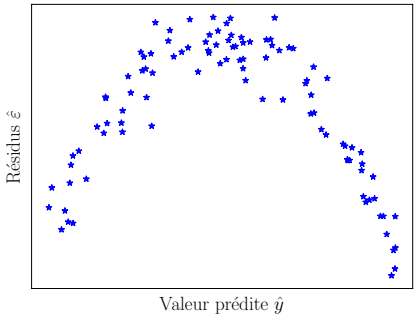


Figure 1: Inédaquation. Solution $\rightarrow$ Rajout d'un régresseur, transformation des entrées

P1 & P2: adéquation et homoscédasticité

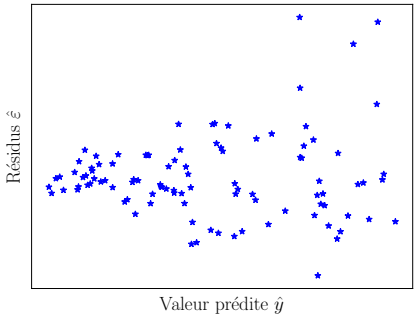


Figure 2: Hétéroscédasticité. Solution $\rightarrow$ transformation de la variable de sortie

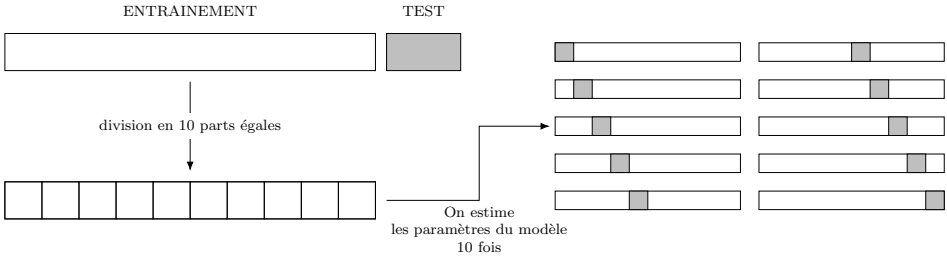
Validation du modèle

Echantillon d'entrainement et de test:
On sépare l'échantillon de données en deux, un pour entrainer le modèle (i.e. résoudre le problème d'optimisation des moindres carrés). L'échantillon restant sera celui de test, permettant de vérifier si le modèle est en adéquation avec les données.

Quel est la limitation de cette méthode ?

Validation croisée

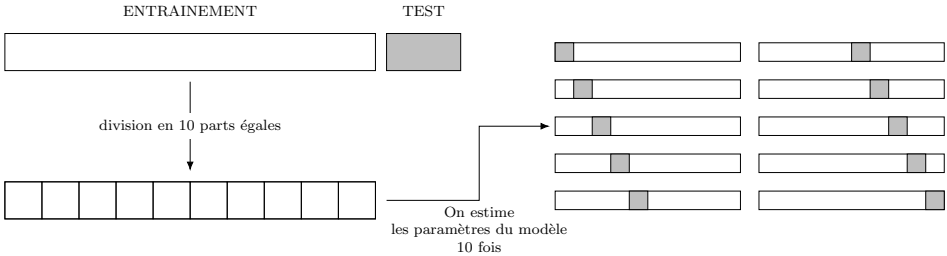
La validation croisée repose sur le même principe que la séparation en base d'entrainement et test, mais en découpant l'échantillon en plusieurs blocs.



Si le nombre de blocs = le nombre de données, on parle de *Leave-One-Out*.

Validation croisée

La validation croisée repose sur le même principe que la séparation en base d'entrainement et test, mais en découpant l'échantillon en plusieurs blocs.



Si le nombre de blocs = le nombre de données, on parle de *Leave-One-Out*.

Limitation: On doit estimer les paramètres autant de fois qu'il y a de blocs, ce qui peut être coûteux en temps de calcul.

Capacité d'approximation

Comment évaluer quantitativement la capacité d'approximation !

↪ **Coefficient de détermination** R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Interprétation: C'est la part de variance expliqué par le modèle.

↪ **Coefficient de prédiction** Q^2. C'est la même formule mais sur un **échantillon de test** (i.e. pas utilisé pour estimer les coefficients du modèle).

Inéquation du modèle

Question: Comment faire pour régler les problèmes d'inédaquation du modèle ?

Inéquation du modèle

Question: Comment faire pour régler les problèmes d'inédaquation du modèle ?

Tentative: augmenter le nombre de variable du modèle ?

Inédaquation du modèle

Testons cette idée avec un modèle polynomial !

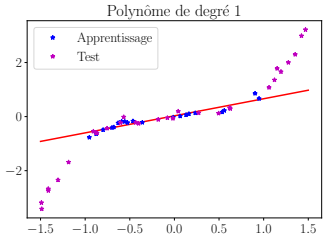
$$y(x) = x^3 + \varepsilon$$

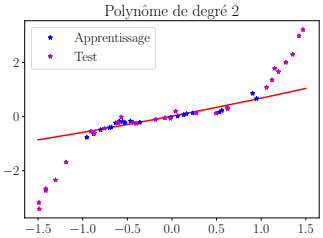
On va choisir comme modèle un polynôme de degré p:

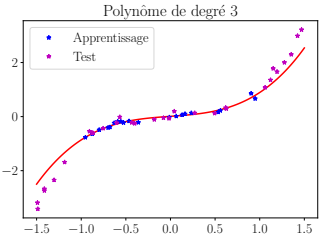
$$\phi(x, \beta) = \sum_{i=1}^{p} \beta_i x^i$$

On remarque que c'est un modèle linéaire (⚠ linéaire en β !). On va voir comment se comporte $\phi(\cdot, \hat{\beta})$ en fonction de p

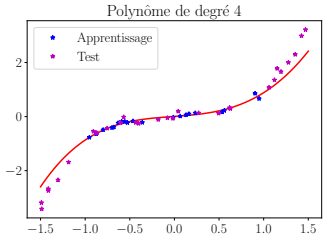
Inéquation du modèle

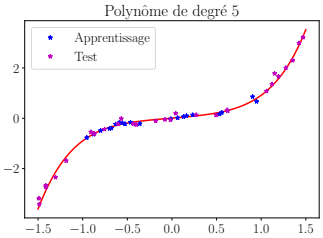


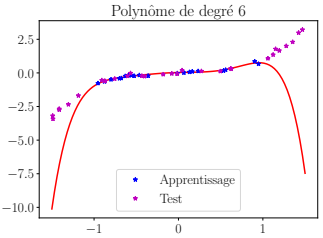




Inéquation du modèle



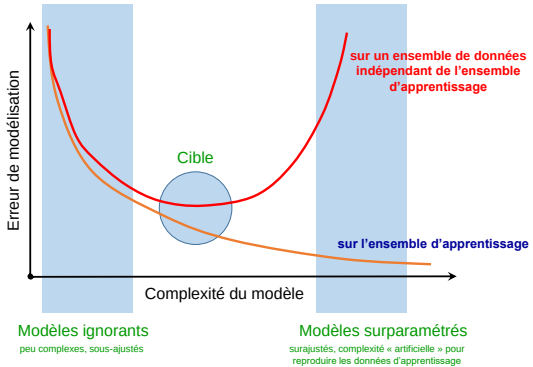




Surapprentissage

En réalité, ajouter plus de variables (et donc plus de paramètres) va améliorer la qualité de prédiction uniquement sur les données observées !

C'est ce qu'on appelle le surapprentissage (*overfit*)



Sélection de modèles

Comment choisir le modèle de dimension "optimal" ?

↪ Critères quantitatifs:

$$\begin{aligned} \mathrm{AIC} &= 2(p+1) - 2\log(\mathcal{L}(\hat{\beta})) \\ \mathrm{BIC} &= n(p+1) - 2\log(\mathcal{L}(\hat{\beta})) \end{aligned}$$

Le modèle optimal minimise ses critères.

↪ Pénalisation: On modifie le critère des moindres carrés dans le but de contraindre l'optimisation des paramètres du modèle.
Exemple:

- La régression Lasso impose d'avoir des paramètres strictement nul. On appelle ça la parcimonie (*sparse*).
- La régression Ridge permet de diminuer la variance de $\hat{\beta}$ par régularisation. Utile quand les régresseurs sont fortement corrélés.

Régression Ridge

C'est une méthode de régularisation, on va pénaliser les moindres carrés de la façon suivante:

$$\begin{aligned}\hat{\beta}_R &= \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \|\mathbf{y} - \boldsymbol{X}\beta\|^2 + \lambda \|\beta\|_2^2 \\ \hat{\beta} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_{p+1})^{-1}\boldsymbol{X}^T\mathbf{y}\end{aligned}$$

On peut remarque que c'est équivalent au problème d'optimisation sous contrainte suivant:

$$\hat{\beta}_R = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \|\mathbf{y} - \boldsymbol{X}\beta\|^2; \ \|\beta\|_2^2 < C$$

On peut ainsi interpréter la régresion Ridge comme une contrainte sur la norme 2 de $\hat{\beta}$, évitant que des composantes prennent des valeurs extrêmes et *de facto* limite la variance des prévisions.

Difficulté: Comment choisir λ ?

Régression Lasso

On minimise une version pénalisé des moindres carrés. On définit $\|\beta\|_1 = \sum_j |\beta_j|$

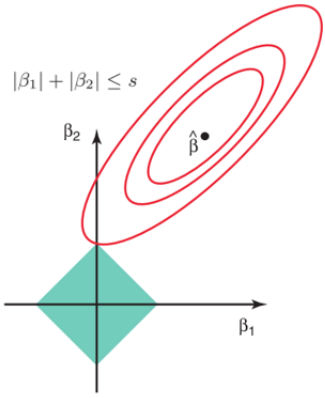
$$\hat{\beta}_L = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \|\mathbf{y} - \boldsymbol{X}\beta\|^2 + \lambda \|\beta\|_1$$

Pas de solution analytique, on doit recourir à des méthodes numériques. On peut remarque que c'est équivalent au problème d'optimisation sous contrainte suivant:

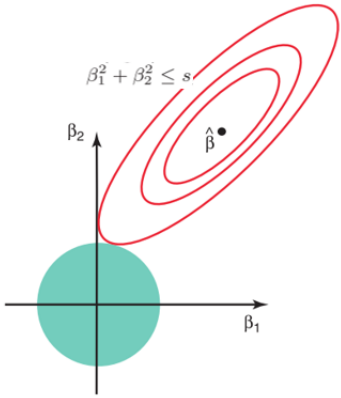
$$\hat{\beta}_L = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \|\mathbf{y} - \boldsymbol{X}\beta\|^2; \ \|\beta\|_1 < C$$

La régression Lasso induit de la parcimonie, c'est à dire que l'on aura $\beta_j = 0$ pour certains indices j.
On fait ainsi naturellement de la sélection de variables ! (Pourquoi à votre avis ?)

Interprétation graphique



Lasso Regression



Ridge Regression

Sommaire

Apprentissage supervisé

Le modèle linéaire permet de faire de l'apprentissage supervisé.

Apprentissage supervisé: On observe une variable Y conjointement avec des données X. Le but de l'apprentissage supervisé est de trouver $\hat{f}$ dans le but de reproduire Y à partir de X:

$$Y = \hat{f}(X) + \varepsilon$$

avec ε l'erreur de prédiction (qu'on cherche à minimiser).

Classification versus régression

Si la sortie Y est à valeurs dans $\mathbb{R}^p$ on parle de régression.

Si la sortie Y est à valeurs dans $\{0, \ldots, K\}$, on parle de classification.

Exemple de régression: Y est une concentration d'une espèce chimique, le prix d'un bien, ...

Exemple de classification: Y est la présence d'un cancer, la détection d'une particule,...

Estimation versus apprentissage

Quel est la différence entre estimer et prédire ?

L'estimation est un principe central de la tradition statistique, elle vise à approcher un *vrai modèle* supposé exister et basé eventuellement sur des théories physiques, biologiques, économique... On utilisant très souvent un cadre probabiliste pour ce genre de problèmes.

Si l'objectif n'est que de prédier, le meilleur modèle d'apprentissage $\hat{f}$ n'est pas forcément celui qui ajusterait le vrai modèle ! La théorie de l'apprentissage est basé sur une notion de qualité de *prédiction*. On choisira des modèles *parcimonieux* (i.e. avec un nombres de paramètres limités) sans trop se préoccuper de l'interprétabilité.