Modèle linéaire

M1 Maths IA

Yannig Goude [1], [2]

[1] EDF R&D, EDF Lab Saclay
[2] Laboratoire de Mathématiques d'Orsay

Janvier 2025

Sommaire

Régression linéaire

Cadre général

Formulation

La régression recouvre plusieurs méthodes d'analyse statistique permettant d'approcher une variable aléatoire Y partir un ensemble d'autres variables aléatoires $X_1, X_2, \ldots, X_p$, regroupées dans un vecteur aléatoire $\boldsymbol{X}$, qui lui sont corrélées

Soient

- $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ une fonction de coût
- $\mathcal{F}$ un espace de fonctions dans lequel la solution est recherchée

L'objectif est de résoudre le problème de minimisation suivant :

$$\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(\boldsymbol{X}))\right]$$

Cadre général

En pratique

- $\mathbb{E}\left[\ell(Y, f(\boldsymbol{X}))\right]$ n'est pas connue
- $\mathcal{F}$ et ℓ peuvent être choisis

$\mathbb{E}\left[\ell(Y, f(\boldsymbol{X}))\right]$ est estimée grâce à un échantillon de réalisations de $(Y, f(\boldsymbol{X}))$

Abus de notation :

- $\boldsymbol{Y} = (y_1, y_2, \ldots y_n)$ est un vecteur de n réalisations de la variable aléatoire Y
- $\mathbf{X}$ est une matrice à n lignes et p colonnes de n réalisations $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots x_{ip})$ du vecteur aléatoire $\boldsymbol{X}$

Cadre général

Approximation

$\mathbb{E}\left[\ell(Y, f(\boldsymbol{X}))\right]$ est approximée par sa moyenne empirique :

$$\mathbb{E}\left[\ell(Y, f(\boldsymbol{X}))\right] \approx \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big)$$

Nous chercherons donc le meilleur modèle $\widehat{f} : \mathbb{R}^p \to \mathbb{R}$ tel que

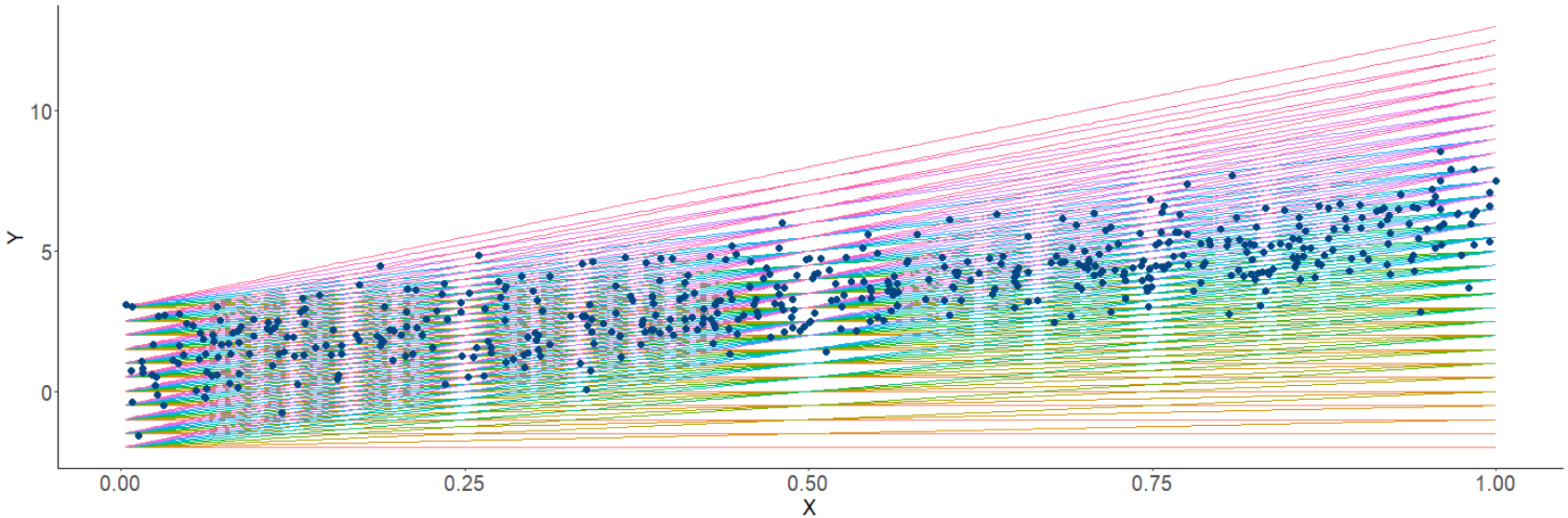
$$\widehat{f} \in \text{argmin}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big)$$

Exemple

Régression linéaire uni-variée

$$\mathcal{F} = \left\{ f_{\alpha,\beta} : x \mapsto \alpha + x\beta \right\}$$

$$\ell : \begin{array}{ccc} \mathbb{R} \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ (y, x) & \longmapsto & (y - x)^2 \end{array}$$

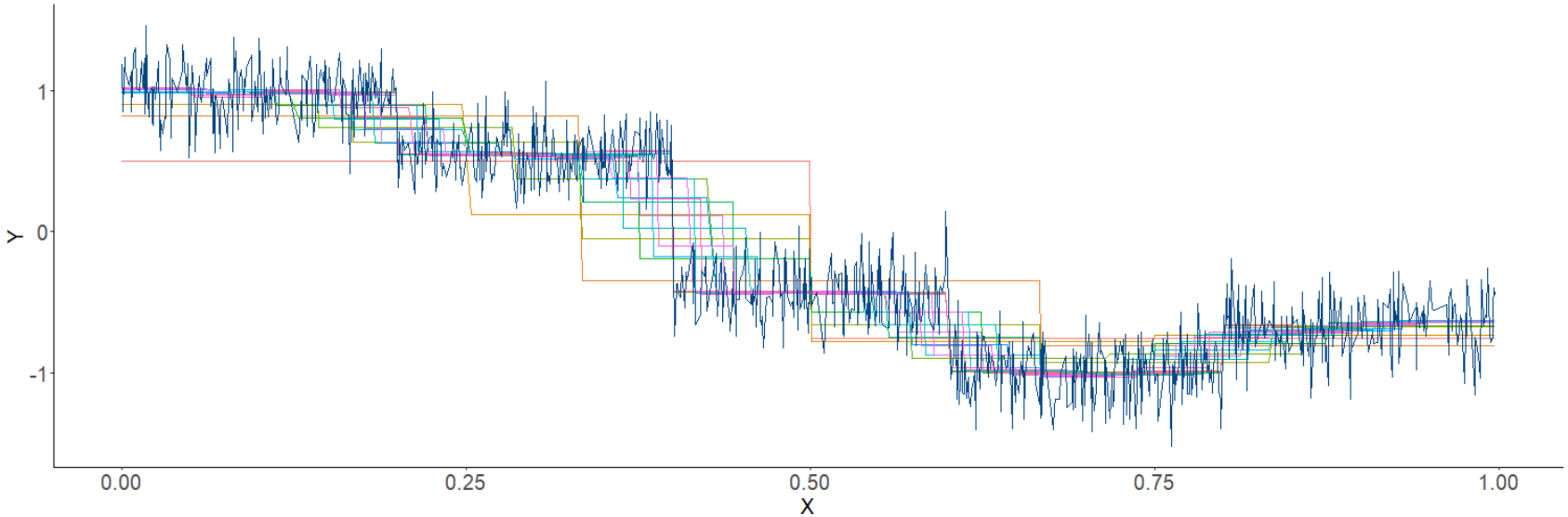


Exemple

Détection de ruptures

$$\mathcal{F} = \left\{ f_{x_0, a_0, \ldots, x_K, a_K} : x \mapsto \sum_{k=1}^{K} a_k \mathbb{1}_{x_{k-1} \leq x < x_k}(x) \right\}$$

$$\ell : \begin{array}{ccc} \mathbb{R} \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ (y, x) & \longmapsto & (y - x)^2 \end{array}$$



Modèle linéaire uni-varié

Formulation

Soient n observations (y_i, x_i), avec $y_i \sim Y_i$ où Y_i est une variable aléatoire

Le modèle linéaire suppose que :

$$y_i = \beta^\star x_i + \varepsilon_i$$

avec $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ indépendants d'espérance nulle et de variance constante

Le processus $(\varepsilon_i)_i$ est un bruit blanc

Modèle linéaire uni-varié

Moindres carrés ordinaires

L'objectif est de trouver β de façon à minimiser les erreurs d'estimation au carré. En dérivant l'erreur quadratique sur l'échantillon constitué des n observations $(y_i, x_i)_{i=1,\dots,n}$

$$\frac{\partial Err(\beta)}{\partial \beta} = \frac{\partial\left(\sum_{i=1}^{n}(y_i - \beta x_i)^2\right)}{\partial \beta} = -\sum_{i=1}^{n} 2x_i(y_i - \beta x_i) = 0$$

L'estimateur des moindres carrés ordinaires est

$$\hat{\beta}^{MCO} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Modèle linéaire uni-varié

Distribution de l'estimateur MCO

Sous l'hypothèse de normalité de la variable aléatoire Y_i :

$$Y_i \sim \mathcal{N}\left(\beta x_i, \sigma^2\right)$$

L'estimateur des moindres carrés ordinaires est sans biais et vérifie

$$\hat{\beta}^{MCO} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Modèle linéaire multi-varié

Formulation

Soient n observations $(y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,p})$, le modèle linéaire suppose que

$$y_i = \beta_1 x_{i,1} + \beta_1 x_{i,2} + \cdots + \beta_1 x_{i,p} + \varepsilon_i$$

Avec

- $x_{i,j} \in \mathbb{R}$ considéré comme déterministe
- $\beta_1, \beta_2, \ldots, \beta_p$ inconnus
- $\mathbb{E}[\varepsilon_i] = 0$ et $\mathrm{Var}(\varepsilon_i) = \sigma^2$

Remarque : une constante peut être inclue ou non dans modèle ($x_{i,1} = 1$)

Modèle linéaire multi-varié

Écriture matricielle

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$ vecteur d'observation et $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n$ vecteur de bruit
- $\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ & \vdots x_{i,j} \vdots & \\ x_{n,1} & \dots & x_{n,p} \end{bmatrix} \in \mathbb{R}^{n \times p}$ matrice de *design* de rang p
- $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^p$ à estimer

Modèle linéaire multi-varié

Moindres carrés ordinaires

Minimisation de la perte quadratique :

$$\widehat{\boldsymbol{\beta}} \in \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \| \boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} \|^2$$

L'estimateur des moindres carrés est obtenu en annulant la différentielle en $\boldsymbol{\beta}$ la perte quadratique, qui est convexe

$$\widehat{\boldsymbol{\beta}}^{MCO} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}$$

Modèle linéaire multi-varié

Biais et variance de l'estimateur MCO

L'estimateur des moindres carrés ordinaire est non biaisé

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}^{MCO}\right] &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{Y}\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right] \\ &= \boldsymbol{\beta}\end{aligned}$$

et de variance

$$\begin{aligned}\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}^{MCO}\right) &= \mathrm{Var}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{Y}\right) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathrm{Var}\left(\boldsymbol{Y}\right)\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\sigma^2\end{aligned}$$

Modèle linéaire multi-varié

MCO et maximum de vraisemblance

La vraisemblance de β au vu des n observations ($\sim$ probabilité d'observer ces observations si celles si sont bien distribuées selon le modèle défini par β) dans le cas où le bruit est gaussien s'écrit

$$L(\mathbf{X}, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{\| \boldsymbol{Y} - \mathbf{X}\beta \|^2}{2\sigma^2} \right)$$

Maximiser la vraisemblance revient à minimiser $\| \boldsymbol{Y} - \mathbf{X}\beta \|^2$ dans le cas gaussien, l'estimateur du maximum de vraisemblance donc égal à l'estimateur des moindres carrés ordinaires

Lorsque les données ne respectent plus l'hypothèse d'indépendance ou de variance constante : $\boldsymbol{Y} \sim (\mathbf{X}\beta, \mathbf{V}\sigma^2)$ où $\mathbf{V}$ est une matrice définie positive,

$$L(\mathbf{X}, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2 |\mathbf{V}|}} \exp\left(- \frac{(\boldsymbol{Y} - \mathbf{X}\beta)^\top \mathbf{V} (\boldsymbol{Y} - \mathbf{X}\beta)}{2\sigma^2} \right)$$

et les deux estimateurs ne sont plus égaux.

Modèle linéaire généralisé

Formulation

Avec les mêmes notation, un modèle linaire généralisé suppose

$$g\big(\mathbb{E}\left[\boldsymbol{Y}\right]\big) = \mathbf{X}\boldsymbol{\beta}$$

où g est une fonction de lien monotone et régulière. Les observations sont toujours supposées indépendantes et suivent une distribution exponentielle.
Une variable aléatoire Y est dans la famille exponentielle si sa densité de probabilité dépend de trois fonctions a, b et c d'un paramètre d'échelle ϕ et d'un paramètre canonique θ de sorte que :

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Modèle linéaire généralisé

Famille exponentielle

	Gaussienne (μ, σ^2)	Poisson (λ)	Binomiale (n, p)	Gamma (α, β)
$\boldsymbol{\theta}$	μ	$\log \lambda$	$\log \frac{p}{1-p}$	$-\frac{\alpha}{\beta}$
$\boldsymbol{\phi}$	σ^2	1	1	$\frac{1}{\alpha}$
$\boldsymbol{a(\phi)}$	ϕ	ϕ	ϕ	ϕ
$\boldsymbol{b(\theta)}$	$\frac{\theta^2}{2}$	e^{θ}	$n \log(1+e^{\theta})$	$-\log -\theta$
$\boldsymbol{c(y,\phi)}$	$\frac{1}{2}\left[\frac{y^2}{\phi} + \log 2\pi\phi\right]$	$-\log y!$	$\log \binom{n}{y}$	$\frac{1}{\phi}\log\left(\frac{y}{\phi}\right) - \log\left(y\Gamma\left(\frac{1}{\phi}\right)\right)$
$\boldsymbol{f(y)}$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	$\frac{\lambda^y e^{-\lambda}}{y!}$	$\binom{n}{y} p^y (1-p)^{n-y}$	$\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$

Exemples :

- Gaussienne : modélisation de la consommation électrique
- Poisson : modélisation d'arrivée/départ à des stations de charge de véhicules électriques

Modèle linéaire généralisé

Maximum de vraisemblance et Algorithme IRLS

Si la variable aléatoire Y est dans la famille exponentielle alors

$$\mathbb{E}[Y] = b'(\theta) \text{ et } \text{Var}(Y) = b''(\theta)a(\phi)$$

Dans le modèle linéaire généralisé, $g\big(\mathbb{E}[\boldsymbol{Y}]\big) = \mathbf{X}\boldsymbol{\beta}$ et la vraisemblance de $\boldsymbol{\beta}$ au vu des n observations s'écrit

$$L(\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_{a_i,b_i,c_i,\theta_i,\phi_i}(y_i)$$

Comme il est alors difficile de maximiser de manière exacte la vraisemblance, la méthode de Newton (méthode numérique avec étape de calcul du gradient et de la Hessienne de la log-vraisemblance) est utilisée pour estimer itérativement $\boldsymbol{\beta}$

À chaque itération k, $\boldsymbol{\beta}^{[k]}$est solution d'un problème des moindres carrés pondérés

$\rightarrow$ Algorithme IRLS : iteratively re-weighted least square (cf. Wood)

Régression linéaire pénalisée

Compromis Biais-Variance

Modèle linéaire pour la prévision

L'hypothèse de modèle linéaire permet d'estimer $\hat{f}(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ à partir d'un échantillon de n observations $(y_i, \boldsymbol{X}_i)_{i=1,\dots,n}$ grâce la méthode des moindres carrés ordinaires (estimateur sans biais et de variance minimale parmi les estimateurs sans biais - Théorème de Gauss-Markov)

Objectif en prévision :

- Prévision de y_{n+1} à partir des variables explicatives $\boldsymbol{X}_{n+1}$

$$\hat{y}_{n+1} = \boldsymbol{X}_{n+1}\hat{\boldsymbol{\beta}}$$

- Observation de y_{n+1} et calcul de l'erreur quadratique (MSE Mean Squared Error)

$$\left(\hat{y}_{n+1} - y_{n+1}\right)^2$$

Compromis Biais-Variance

Erreur de prévision

Sous les hypothèses du modèle linéaire, calcul de l'espérance de l'erreur quadratique :

$$\begin{aligned}
\mathbb{E}\left[\left(y_{n+1} - \hat{y}_{n+1}\right)^2\right] &= \mathbb{E}\left[\left(\boldsymbol{X}_{n+1}\boldsymbol{\beta} + \varepsilon_{n+1} - \boldsymbol{X}_{n+1}\hat{\boldsymbol{\beta}}\right)^2\right] \\
&= \sigma^2 + \mathbb{E}\left[\left(\boldsymbol{X}_{n+1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right)^2\right] \\
&= \sigma^2 + \boldsymbol{X}_{n+1}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{X}_{n+1} + \left(\mathbb{E}\left[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right]\right)^2
\end{aligned}$$

Erreur en prévision = Erreur irréductible + Variance + Biais^2

Compromis Biais-Variance

Illustration

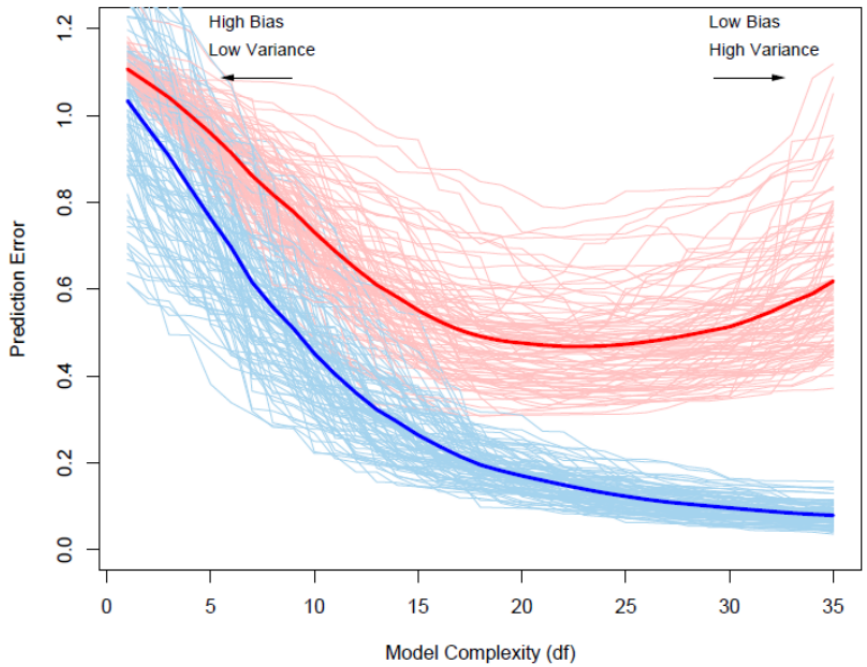


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

The elements of statistical learning : Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer series in statistics - 2001

Régression Ridge

Motivation

Cas extrême :

- Modèle uni-varié : $\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \varepsilon$
- Ajout d'une variable explicative : $\boldsymbol{X}_2 = \boldsymbol{X}_1 + \text{bruit}$

$$\forall a \in \mathbb{R}, \quad \boldsymbol{\beta}_a = \begin{bmatrix} (a+1)\beta_1 \\ -a\beta_1 \end{bmatrix}$$

est un estimateur qui donne une prévision non biaisée

$$\mathbb{E}\left[\hat{\boldsymbol{Y}}\right] = \mathbb{E}\left[(a+1)\boldsymbol{X}_1\beta_1 - a\boldsymbol{X}_2\beta_1\right] = \boldsymbol{X}_1\beta_1 = \mathbb{E}\left[\boldsymbol{Y}\right]$$

mais de variance

$$\begin{aligned} \text{Var}\left(\hat{\boldsymbol{Y}}\right) &= \mathbb{E}\left[\left((a+1)\boldsymbol{X}_1\beta_1 - a\boldsymbol{X}_2\beta_2 - \boldsymbol{X}_1\beta_1\right)^2\right] \\ &= a^2\beta_1^2\text{Var}\left(\text{bruit}\right) \end{aligned}$$

Régression Ridge

Motivation

Si les coefficients de $\boldsymbol{\beta}$ ne sont pas contraints, ils peuvent

- exploser
- être soumis à une très grande variance

S'il existe des variables corrélées (β_1 et β_2 dans l'exemple précédent) dans un modèle de régression linéaire, leurs coefficients peuvent être mal déterminés et présenter une variance élevée : Un coefficient extrêmement élevé sur une variable peut être annulé par un coefficient négatif également élevé sur son cousin corrélé

$\rightarrow$ Imposer une contrainte de taille sur les coefficients

Régression Ridge

Pénalisation

Pour éviter ces écueils, les coefficients de $\boldsymbol{\beta}$ doivent être contraints, et le problème de régression Ridge devient

$$\text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{avec} \quad \|\boldsymbol{\beta}\|^2 \leq \text{constante}$$

Ce problème équivaut à résoudre, avec $\lambda > 0$

$$\text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

Régression Ridge

Estimateur, Biais et Variance

L'expression explicite de l'estimateur Ridge s'obtient par différenciation

$$\frac{\partial\left(\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|^2\right)}{\partial\boldsymbol{\beta}}=2\mathbf{X}^\top\left(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\right)+2\lambda\boldsymbol{\beta}$$

$$\rightarrow \hat{\boldsymbol{\beta}}_\lambda=\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\boldsymbol{Y}$$

L'estimateur Ridge est biaisé :

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}_\lambda\right]=\mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\beta}+\varepsilon\right)\right]=\boldsymbol{\beta}-\lambda\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}\boldsymbol{\beta}$$

et de variance

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_\lambda\right)=\sigma^2\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}$$

Régression Lasso

Motivation et Formulation

La régression Lasso (*least absolute shrinkage and selection operator*) a été introduite dans une optique de sélection de variables et sous l'hypothèse que le vecteur $\boldsymbol{\beta}$ est parcimonieux (grand nombre de ses coefficients sont nuls)

$$\text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{avec} \quad \|\boldsymbol{\beta}\|_1 \leq \text{constante}$$

Ce problème équivaut à résoudre, avec $\lambda > 0$

$$\text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Remarque : la norme 1 est préférée à la norme 0 pour garder un problème sous-différenciel (cf. Introduction to High-Dimensional Statistics, Christophe Giraud, 2014)

Régression Ridge et Lasso

Illustration

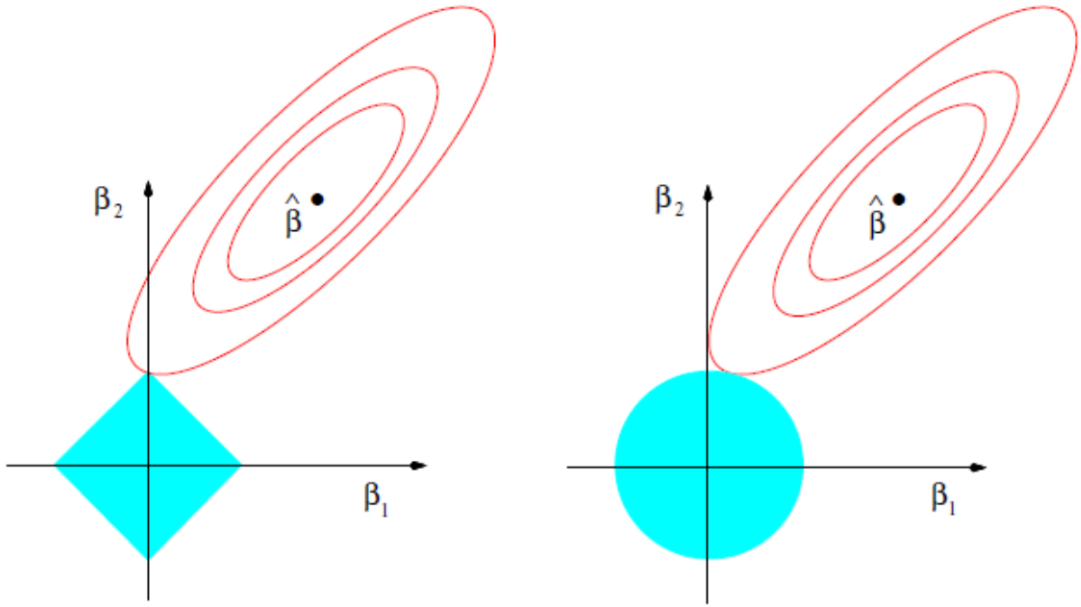


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

The elements of statistical learning : Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer series in statistics - 2001

Choix de la pénalité

Comportements extrêmes

Les estimateurs Ridge et Lasso dépendent fortement du paramètre λ

- Chaque λ donne une unique solution
- λ est un paramètre de régularisation
  - $\lambda = 0 : \hat{\boldsymbol{\beta}}_{\lambda}^{Ridge} = \hat{\boldsymbol{\beta}}_{\lambda}^{Lasso} = \hat{\boldsymbol{\beta}}^{MCO}$
  - $\lambda \to \infty : \hat{\boldsymbol{\beta}}_{\lambda}^{Ridge} = \hat{\boldsymbol{\beta}}_{\lambda}^{Lasso} = \mathbf{0}$

Le paramètre λ gère le compromis biais variance :

- $\lambda = 0 : \mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right] = \mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Lasso}\right] = \mathbb{E}\left[\hat{\boldsymbol{\beta}}^{MCO}\right] = \boldsymbol{\beta}$ mais la variance des estimateurs peut exploser (cf. exemple motivation Ridge)
- $\lambda \to \infty : \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right) = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\lambda}^{Lasso}\right) = \mathbf{0}$ mais les estimateurs sont forcement biaisés (de biais $-\boldsymbol{\beta}$)

Comment choisir λ ?

Choix de la pénalité

Sélection de modèle

Choisir le meilleur λ de façon à avoir des performances optimales en prévision fait parti d'une gamme de problème plus large : la sélection de modèles

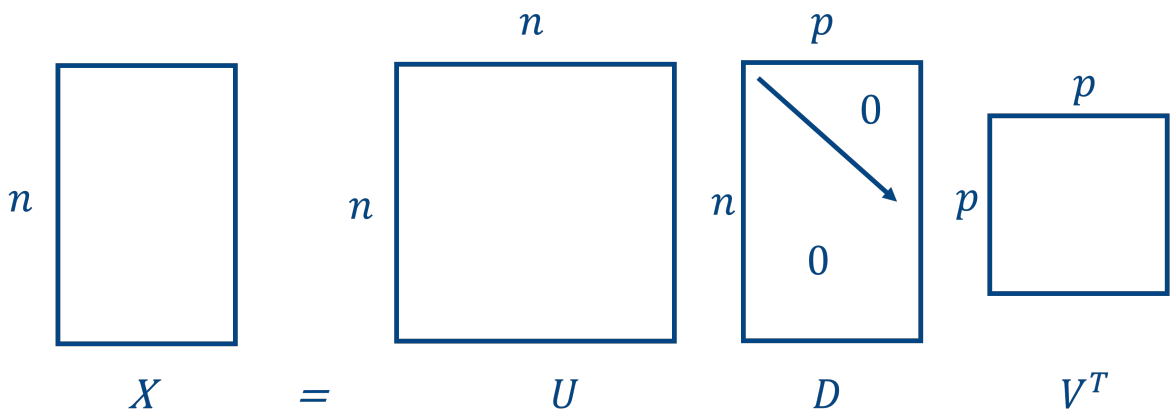
Afin d'avoir un modèle performant en prévision, il convient de construire un ensemble de données dites d'estimation (ou d'apprentissage) et d'un second ensemble de données dites de validation (ou de test) sur lesquelles le modèle se doit d'être performant

Des techniques de validation croisée (qui consiste à tirer plusieurs ensembles de validation d'une même base de données pour obtenir une estimation plus robuste) permettent de trouver les paramètres de régularisation optimaux

Avant d'expliciter le critère de validation pour le choix du paramètre de régularisation de la régression Ridge, nous allons définir les degrés de liberté effectifs d'un modèle (qui nécessite un rappel sur la décomposition en valeurs singulières)

Critère de validation croisée

Décomposition en valeurs singulières (SVD) et nouvelle expression de $\hat{\beta}_\lambda^{Ridge}$



Les matrices U et V sont orthogonales ($\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$) et D contient dans ses coefficients diagonaux les valeurs singulières de $\mathbf{X}$ (elles correspondent aux racines des valeurs propres).

$$\hat{\beta}_\lambda^{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{Y} = \mathbf{V}\mathrm{Diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\mathbf{U}^\top\boldsymbol{Y}$$

Critère de validation croisée

Matrice de lissage et *Effective Degrees of Freedom*

Une matrice de lissage $\mathbf{A}$ est un opérateur linéaire tel que :

$$\hat{\boldsymbol{Y}} = \mathbf{A}\boldsymbol{Y}$$

- Moindres Carrés Ordinaires : $\mathbf{A}^{MCO} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$

  Notons que la trace de $\mathbf{A}^{MCO}$, $\mathrm{Tr}\left(\mathbf{A}^{MCO}\right) = p$,le nombre de paramètres à estimés. Par analogie, les *Effective Degrees of Freedom* d'une matrice de lissage $\mathbf{A}$ sont définis par

  $$\mathrm{df}\left(\mathbf{A}\right) = \mathrm{Tr}\left(\mathbf{A}\right)$$

- Régression Ridge : $\mathbf{A}_\lambda^{Ridge} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top$

$$\mathrm{df}\left(\mathbf{A}_\lambda^{Ridge}\right) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Critère de validation croisée

Principe

$\forall i = 1, \ldots, n$

- Retirer l'observation $(y_i, \boldsymbol{X}_i)$ de l'échantillon
- Estimer $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ à l'aide de toutes les autres données
- Mesurer l'erreur de prévision $(y_i - \hat{\boldsymbol{\beta}}_\lambda^{-i} X_i)^2$

Le critère de validation croisée s'écrit

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\beta}}_\lambda^{-i} X_i)^2$$

Il est possible de montrer qu'en espérance, le critère de validation croisée est égal à l'erreur de prévision ; il convient donc de le minimiser

Critère de validation croisée

Généralisation

Un tel critère prend du temps à être calculé (n estimateurs pour chaque λ) !

Mais, pour la régression Ridge, il est possible de montrer que

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_\lambda^{-i} X_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\beta}_\lambda X_i)^2}{\left(1 - \mathbf{A}_{\lambda_{i,i}}\right)^2}$$

Il n'y a donc plus qu'un seul estimateur à calculer !

L'approximation $\mathbf{A}_{\lambda_{i,i}} \approx \frac{\mathrm{Tr}}{(\mathrm{A}_\lambda)n}$ permet de proposer le critè de validation croisée généralisé :

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i - \hat{\beta}_\lambda \boldsymbol{X}_i\right)^2}{\left(1 - \frac{\mathrm{df}\left(\mathbf{A}_\lambda\right)}{n}\right)^2}$$

C'est ce dernier qui est généralement utilisé par les algorithmes pour choisir le meilleur paramètre de régularisation

Régression quantile

Régression quantile

Motivation

Alors que la méthode des moindres carrés fournit une estimation de l'espérance (conditionnellement aux variables explicative) de la variable réponse, la régression quantile cherche à approcher la médiane ou d'autres quantiles de la variable réponse

Elle s'avère utile pour prévoir notamment des seuils et lorsque plusieurs régressions sont réalisées, il est possible d'avoir une bonne idée de la distribution générale de la variable réponse

La régression quantile est moins sensible aux points aberrants. Elle peut être définie comme une régression avec une norme 1

Régression quantile

Formulation

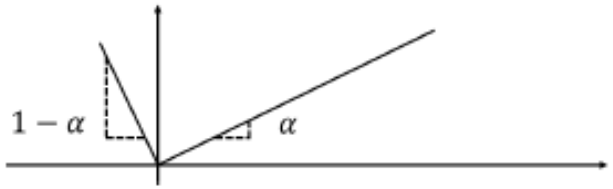
Le quantile q_α de la variable aléatoire Y de densité f et de fonction de répartition F vérifie

$$F(q_\alpha) = \int_{-\infty}^{q_\alpha} f(y)\mathrm{d}y = \mathbb{P}(Y \leq q_\alpha) = \alpha$$

En introduisant la pinball loss

$$\ell_\alpha(y-q) = \alpha|y-q|^+ + (1-\alpha)|y-q|^-,$$

$$\text{avec} \quad |x|^+ = \max(x,0) \quad \text{et} \quad |x|^- = \max(-x,0)$$



Le quantile q_α minimise la quantité $\mathbb{E}[\ell_\alpha(Y-q)]$

Régression quantile

Preuve

Résolvons le problème de minimisation convexe

$$\text{argmin}_q \, \mathbb{E}\left[\ell_\alpha(Y-q)\right]$$

par différentiation :

$$\begin{aligned}
0 &= \mathbb{E}\left[\frac{\partial \ell_\alpha(Y-q)}{\partial q}\right] \\
&= \int_{-\infty}^{+\infty} \frac{\partial \ell_\alpha(y-q)}{\partial q} f(y)\mathrm{d}y \\
&= -(1-\alpha)\int_{-\infty}^{q} f(y)\mathrm{d}y + \alpha \int_{q}^{+\infty} f(y)\mathrm{d}y \\
&= (\alpha-1)F(q) + \alpha(1-F(q)) = \alpha - F(q)
\end{aligned}$$

La solution vérifie donc

$$F(\hat{q}) = \alpha$$

Régression quantile

Estimateur et Algorithme

Lorsque l'on dispose de l'échantillon $(y_1, \ldots, y_n)$, l'estimateur de la régression quantile est obtenu en minimisant le critère

$$\hat{\boldsymbol{\beta}}^{\alpha} \in \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell_{\alpha}(y_i, \boldsymbol{X}_i \boldsymbol{\beta})$$

Il est possible d'utiliser une méthode de descente de gradient puisque la fonction à minimiser est presque partout dérivable. L"algorithme Iteratively reweighted least squares permet aussi de calculer l'estimateur (plusieurs implémentations existent dans différents packages R ou python)

Références

- Wood, Simon N. Generalized additive models: an introduction with R. chapman and hall/CRC, 2006.
- Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.
- Giraud, Christophe. Introduction to high-dimensional statistics. Chapman and Hall/CRC, 2021.