Régression linéaire

Extension à la régression logistique pour la classification

Machine Learning

Cédric RICHARD

Université Nice Sophia Antipolis

Régression linéaire

Exemple de modèle linéaire simple

Problème : On s'intéresse à la concentration d'ozone O_3 dans l'air.

On cherche à savoir s'il est possible d'expliquer la concentration maximale d'ozone de la journée par la température T à midi.

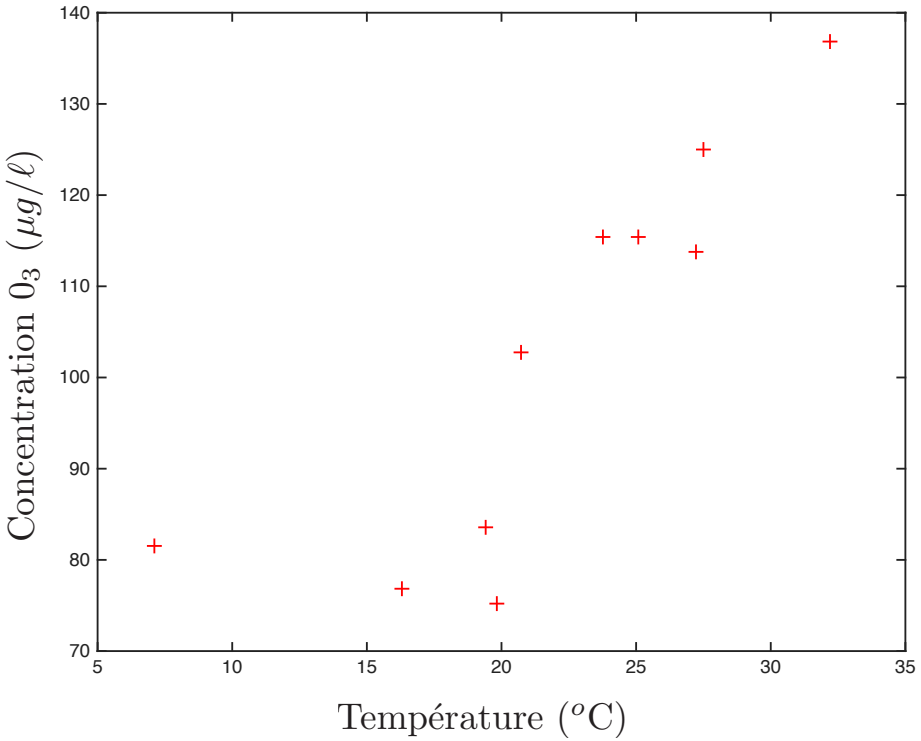
Objectif : Le but de la régression est :

- Ajuster les paramètres d'un modèle pour expliquer O_3 à partir de T
- Prédire O_3 à partir de nouvelles valeurs de T

T	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

Régression linéaire

Exemple de modèle linéaire simple



Régression linéaire

Exemple de modèle linéaire simple

Problème : Le botaniste Joseph D. Hooker a mesuré en 1849, en Himalaya, la pression atmosphérique p_i et la température d'ébullition de l'eau t_i.

Modèle : Selon les lois de la physique, $y_i = \ln p_i$ devrait être en première approximation proportionnel à t_i. On pose donc :

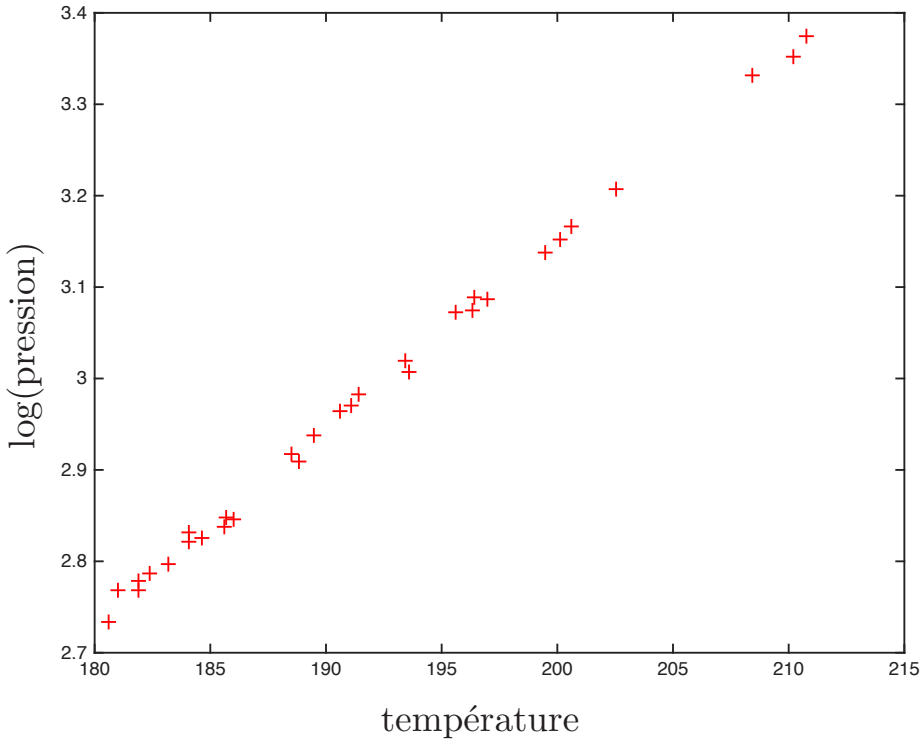
$$y_i = \beta_1 + \beta_2 t_i + \varepsilon_i$$

où ε_i est une erreur d'approximation de moyenne nulle et de variance σ^2.

Objectif : Estimer $\boldsymbol{\beta} = [\beta_1\, \beta_2]^\top$ permettant d'expliquer y_i à partir de t_i.

Régression linéaire

Exemple de modèle linéaire simple



Régression linéaire

Exemple de modèle linéaire multiple

Fonction de production : La fonction de Cobb-Douglas (1928) est largement utilisée en économie comme modèle de fonction de production. Elle exprime le niveau de production p_i d'un bien en fonction du capital utilisé k_i et de la quantité de travail t_i :

$$p_i = \alpha_1 \cdot k_i^{\alpha_2} \cdot t_i^{\alpha_3}$$

où α_1, α_2 et α_3 sont déterminés par la technologie.

Modèle : Le modèle de Cobb-Douglas peut être linéarisé ainsi :

$$\ln p_i = \ln \alpha_1 + \alpha_2 \ln k_i + \alpha_3 \ln t_i + \varepsilon_i$$

où ε_i est une erreur d'approximation de moyenne nulle et de variance σ^2.

Objectif : Estimer $\boldsymbol{\beta} = [\ln(\alpha_1)\, \alpha_2\, \alpha_3]^\top$.

Régression linéaire

Exemple de modèle linéaire multiple

Données : Cobb et Douglas (1928) disposent des données suivantes :

Année	p_i	k_i	t_i	Année	p_i	k_i	t_i	Année	p_i	k_i	t_i
1899	100	100	100	1907	151	176	138	1915	189	266	154
1900	101	107	105	1908	126	185	121	1916	225	298	182
1901	112	114	110	1909	155	198	140	1917	227	335	196
1902	122	122	118	1910	159	208	144	1918	223	366	200
1903	124	131	123	1911	153	216	145	1919	218	387	193
1904	122	138	116	1912	177	226	152	1920	231	407	193
1905	143	149	125	1913	184	236	154	1921	179	417	147
1906	152	163	133	1914	169	244	149	1922	240	431	161

Résultat : Dans le cas de rendements d'échelle constants ($\alpha_2 + \alpha_3 = 1$), Cobb et Douglas trouvent :

$$\alpha_2 = \frac{1}{4} \qquad \alpha_2 = \frac{3}{4}$$

Régression linéaire

Notations

▷ Les données consistent en des variables observées y_i (réponses) et des variables explicatives (ou régresseurs) $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]^\top$, avec $i = 1, \ldots n$. Chaque paire $(y_i, \boldsymbol{x}_i)$ représente une expérience (un individu).

▷ On présuppose l'existence d'une relation de la forme :

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \ldots + \beta_p \, x_{ip} + \varepsilon_i$$

où ε_i est une erreur d'approximation.

▷ On suppose que $\boldsymbol{\varepsilon} = [\varepsilon_1 \ldots, \varepsilon_n]^\top$ est un n-échantillon d'espérance nulle. Souvent, il est supposé distribué selon une loi Gaussienne $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Régression linéaire

Notations

▷ On arrange les n individus de p variables explicatives ainsi :

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \\ \vdots \\ \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & & \cdots & & x_{1p} \\ 1 & x_{21} & x_{22} & & & & \\ & & & \ddots & & & \\ \vdots & & & & x_{ij} & & \vdots \\ & & & & & \ddots & \\ 1 & x_{n1} & & & \cdots & & x_{np} \end{pmatrix}$$

On convient de mettre le régresseur constant (associé au paramètre β_0), s'il y a lieu, dans la première colonne de $\boldsymbol{X}$.

▷ Pour simplifier les notations, et éviter de devoir distinguer les cas avec et sans composante continue β_0, on supposera que $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ et $\boldsymbol{\beta} \in \mathbb{R}^p$.

Régression linéaire

Modèle linéaire

Définition 1. *Un modèle linéaire se définit par une équation de la forme :*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où :

- ▷ $\boldsymbol{y} \in \mathbb{R}^n$ *est le vecteur des observations*
- ▷ $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ *est la matrice des variables explicatives (régresseurs)*
- ▷ $\boldsymbol{\beta} \in \mathbb{R}^p$ *le vecteur des coefficients à estimer.*

Hypothèses. On suppose que la matrice $\boldsymbol{X}$ est de rang plein p, et que le vecteur de bruit $\boldsymbol{\varepsilon}$ est un n-échantillon centré de variance σ^2.

Régression linéaire

Estimateur des moindres carrés ordinaires

Objectif. Les points $(y_i, \boldsymbol{x}_i)$ étant observés, il s'agit d'estimer la fonction affine définie par $\hat{f} : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}$ minimisant l'erreur quadratique moyenne :

$$\hat{f} \triangleq \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i)\right)^2$$

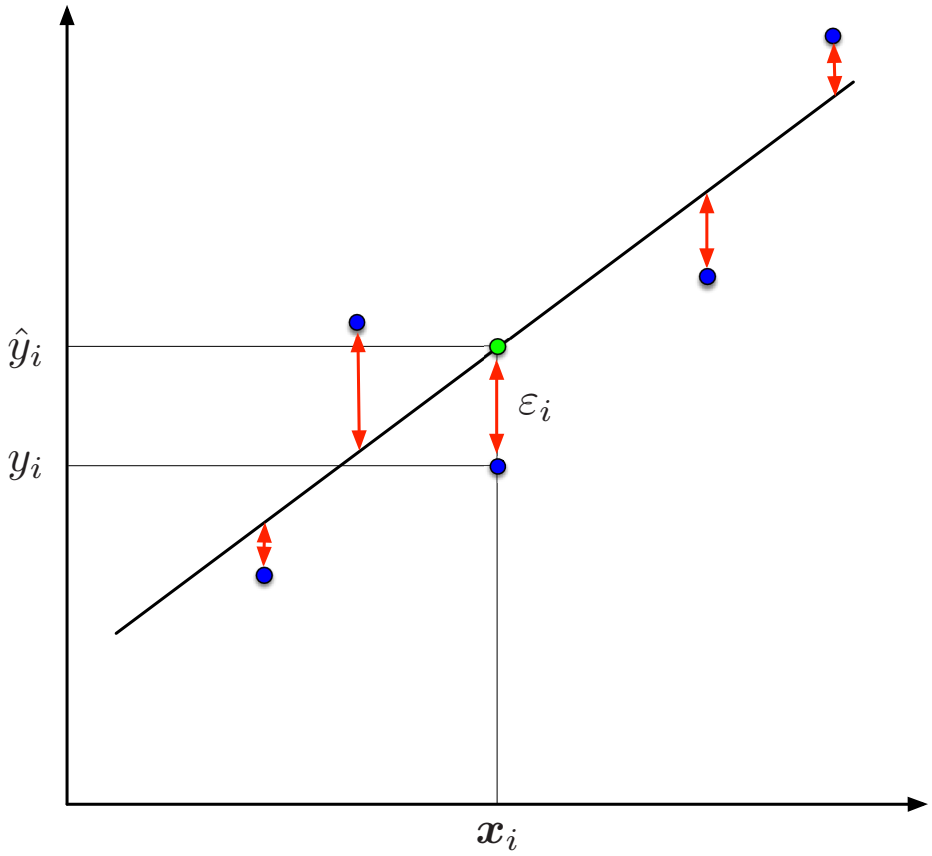
Définition 2 (Estimateur des moindres carrés ordinaire (OLS))**.** *On appelle estimateur des moindres carrés ordinaire, le minimiseur $\widehat{\boldsymbol{\beta}}$ du risque empirique :*

$$\begin{aligned}\hat{f} &\triangleq \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 \\ &\triangleq \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2\end{aligned}$$

où $\|\cdot\|^2$ désigne la norme euclidienne telle que $\|\boldsymbol{\varepsilon}\|^2 \triangleq \sum_{i=1}^{n} \varepsilon_i^2$.

Régression linéaire

Estimateur des moindres carrés ordinaires



Régression linéaire

Estimateur des moindres carrés ordinaires

Proposition 1 (Estimateur des moindres carrés ordinaire (OLS))**.** *Sous les hypothèses et les notations de la Définition 1, l'OLS existe et est unique. Il s'écrit :*

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Démonstration :

On considère le problème à résoudre :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) \triangleq \frac{1}{n}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right\|^2$$

Une condition nécessaire d'optimalité est : $\nabla J(\boldsymbol{\beta}) = 0$. Or :

$$\nabla J(\boldsymbol{\beta}) = -\frac{2}{n}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$

On note que $\boldsymbol{X}^\top\boldsymbol{X} \in \mathbb{R}^{p\times p}$ est inversible puisque $\boldsymbol{X}$ est de rang p par hypothèse. Afin que $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ soit optimum au sens de $J(\boldsymbol{\beta})$, il doit donc nécessairement vérifier :

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Cette condition est suffisante car $J(\boldsymbol{\beta})$ est une fonction strictement convexe. En effet, sa matrice Hessienne $\nabla^2 J(\boldsymbol{\beta}) \triangleq \frac{2}{n}(\boldsymbol{X}^\top\boldsymbol{X})$ est définie positive (strictement).

Régression linéaire

Interprétation géométrique

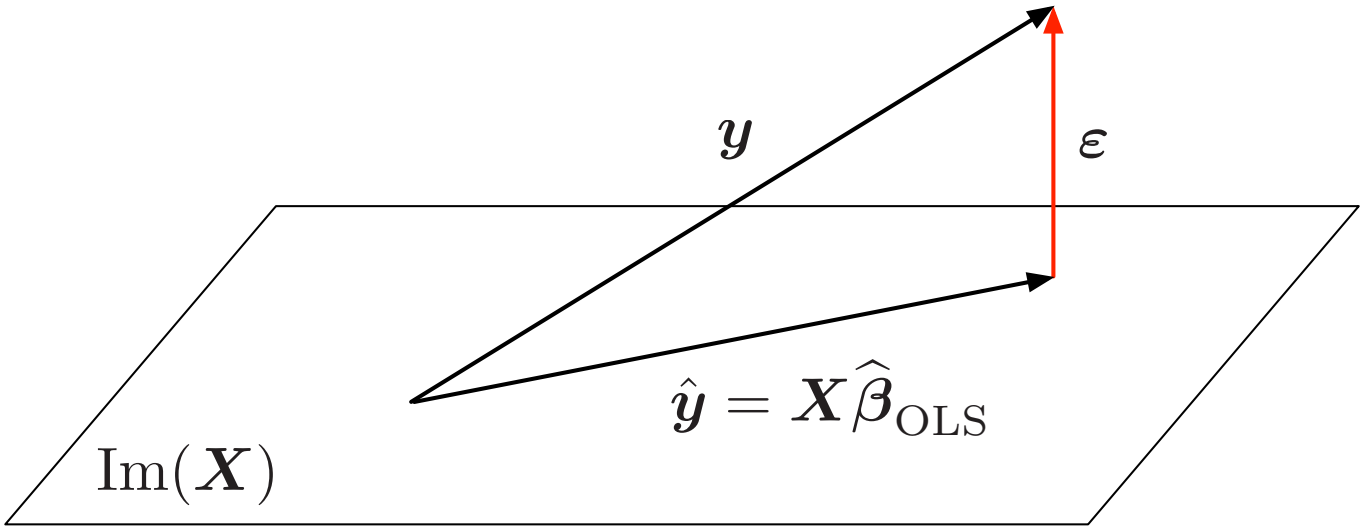
Le modèle linéaire vise à représenter $\boldsymbol{y}$ par une combinaison linéaire $\boldsymbol{X}\boldsymbol{\beta}$ des colonnes de $\boldsymbol{X}$. Celles-ci constituent une famille libre de $\mathbb{R}^n$ puisque $\text{rang}(\boldsymbol{X}) = p$.

En minimisant $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, on recherche donc l'élément de $\text{Im}(\boldsymbol{X})$ le plus proche de $\boldsymbol{y}$ au sens de la distance euclidienne. Il s'agit de la projection orthogonale de $\boldsymbol{y}$ sur $\text{Im}(\boldsymbol{X})$, notée $\hat{\boldsymbol{y}} \triangleq \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.

Notons que $\boldsymbol{P}_X \triangleq (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ est la matrice de projection orthogonale sur le sous-espace $\text{Im}(\boldsymbol{X})$.

Régression linéaire

Interprétation géométrique



Régression linéaire

Propriétés

Propriété 1. *Soit* $\boldsymbol{\varepsilon} \triangleq \boldsymbol{y} - \hat{\boldsymbol{y}}$. *On a :*

$$\boldsymbol{\varepsilon} \perp \hat{\boldsymbol{y}}$$

Démonstration :

$$\begin{aligned}
\boldsymbol{\varepsilon}^\top \hat{\boldsymbol{y}} &= (\boldsymbol{y} - \boldsymbol{P}_X \boldsymbol{y})^\top \boldsymbol{P}_X \boldsymbol{y} \\
&= \|\boldsymbol{P}_X \boldsymbol{y}\|^2 - \|\boldsymbol{P}_X \boldsymbol{y}\|^2 \\
&= 0
\end{aligned}$$

Régression linéaire

Mesures de performance du modèle linéaire

L'erreur quadratique moyenne $\frac{1}{n}\|\boldsymbol{\varepsilon}\|^2$ permet de caractériser la qualité d'un modèle linéaire, mais ce n'est pas une grandeur normalisée.

Il est préférable de recourir au coefficient de détermination, défini comme le rapport de la variance expliquée sur la variance totale :

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

- ▷ $R^2 = 1$: le modèle linéaire est parfait, les points $(\boldsymbol{x}_i, y_i)$ sont alignées
- ▷ $R^2 \approx 0$: le modèle linéaire n'est pas approprié

Propriété 2. *On a :*

$$R^2 = \rho_{y\hat{y}}$$

où ρ désigne le coefficient de corrélation

Régression linéaire

Propriétés

Propriété 3. *L'estimateur $\widehat{\boldsymbol{\beta}}_{OLS}$ est sans biais, $\mathbb{E}\{\widehat{\boldsymbol{\beta}}_{OLS}\} = \boldsymbol{\beta}$, et de covariance :*

$$Cov\{\widehat{\boldsymbol{\beta}}_{OLS}\} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

Démonstration :

$$\begin{aligned}\mathbb{E}\{\widehat{\boldsymbol{\beta}}_{\text{OLS}}\} &= \mathbb{E}\left\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\right\}\\ &= \mathbb{E}\left\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\}\\ &= \boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\text{Cov}\{\widehat{\boldsymbol{\beta}}_{\text{OLS}}\} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbb{E}\left\{\boldsymbol{y}\boldsymbol{y}^\top\right\}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{aligned}$$

Régression linéaire

Propriétés

Pour caractériser l'écart entre $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ et $\boldsymbol{\beta}$, on étudie l'écart quadratique suivant :

$$\mathrm{MSD}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \triangleq \mathbb{E}\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|^2\}$$

Propriété 4.

$$MSD(\widehat{\boldsymbol{\beta}}_{OLS}) = \sigma^2 \, trace\,(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

Démonstration :

$$\begin{aligned}
\mathrm{MSD}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) &\triangleq \mathbb{E}\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|^2\} \\
&= \mathbb{E}\{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\right)^\top \left(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\right)\} \\
&= \mathrm{trace}\Big\{\mathbb{E}\{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\right)^\top\}\Big\} \\
&= \mathrm{trace}\Big\{\mathrm{Cov}\{\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\}\Big\}
\end{aligned}$$

Régression linéaire

Propriétés

La Propriété 4 est peu pratique pour caractériser la précision de $\widehat{\boldsymbol{\beta}}$ car elle fait intervenir la variance du bruit σ^2, inconnue dans un contexte applicatif.

Un estimateur naturel est de considérer l'erreur moyenne :

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)^2$$

Propriété 5. *Dans le cas Gaussien, les estimateurs par maximum de vraisemblance de $\boldsymbol{\beta}$ et σ vérifient :*

$$\widehat{\boldsymbol{\beta}}_{MV} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{y} \quad et \quad \widehat{\sigma}^2_{MV} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n}$$

Régression linéaire

Identifiabilité

Si $\boldsymbol{X}$ n'est pas de rang maximum p, la matrice $\boldsymbol{X}^\top \boldsymbol{X}$ n'est pas inversible et le problème admet plusieurs solutions qui minimisent le risque empirique.

$$\Rightarrow \text{ le problème est mal posé ou non-identifiable.}$$

Plusieurs solutions permettent de traiter ce problème :

- **Par sélection de variables :**
  Cette stratégie consiste à réduire la dimension de l'espace des solutions par sélection des colonnes de $\boldsymbol{X}$ de sorte que celle-ci soit de rang maximum.

- **Par régularisation du problème :**
  Cette stratégie consiste à régulariser le problème de minimisation du risque empirique de sorte que celui-ci admette une solution unique.

Régression Ridge

Problème et solution

La méthode de régression régularisée la plus couramment utilisée est certainement la régression Ridge. Elle vise à contraindre l'amplitude des composantes de $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) \triangleq \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \text{ sous la contrainte } \|\boldsymbol{\beta}\|^2 \leq \tau, \quad \tau > 0$$

Ce problème peut être formulé de manière équivalente (Lagrangien) ainsi :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) \triangleq \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2, \quad \lambda > 0$$

où λ est lié à τ...

Régression Ridge

Problème et solution

Le problème de régression Ridge s'écrit :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) \triangleq \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2, \quad \lambda > 0$$

En suivant le même cheminement que précédemment, on trouve :

Proposition 2. *La solution du problème de régression Ridge existe et est unique pour tout* $\lambda > 0$. *Elle s'écrit :*

$$\widehat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Remarque :

- ▷ Il n'y a plus de problème d'inversion relativement au rang de $\boldsymbol{X}^\top\boldsymbol{X}$ puisque la matrice $\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{I}$ est de rang d.
- ▷ Le choix du paramètre λ est essentiel en pratique (voir ci-après).

Régression Ridge

Propriétés

Propriété 6. *L'estimateur* $\widehat{\boldsymbol{\beta}}_{ridge}$ *vérifie :*

$$\mathbb{E}\{\widehat{\boldsymbol{\beta}}_{ridge}\} = (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{\beta}$$

$$Cov\{\widehat{\boldsymbol{\beta}}_{ridge}\} = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{X})(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1}$$

Remarque :

- ▷ L'estimateur Ridge est biaisé contrairement à l'estimateur OLS, ce qui constitue un inconvénient.
- ▷ La covariance de l'estimateur Ridge ne fait pas intervenir l'inverse de $\boldsymbol{X}^\top \boldsymbol{X}$, mais celui de $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ qui est mieux conditionné.
- ▷ On note que $\lim_{\lambda \to \infty} \mathrm{Cov}\{\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}\} = \mathbf{0}$

Régression Ridge

Propriétés

Démonstration de la Propriété 5 :

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{ridge}} &= (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) \, \widehat{\boldsymbol{\beta}}_{\text{OLS}}
\end{aligned}$$

$$\begin{aligned}
\text{Cov}\{\widehat{\boldsymbol{\beta}}_{\text{ridge}}\} &= \text{Cov}\Big\{ (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) \, \widehat{\boldsymbol{\beta}}_{\text{OLS}} \Big\} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) \, \text{Cov}\{\widehat{\boldsymbol{\beta}}_{\text{OLS}}\} \, (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} \\
&= \sigma^2 \, (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1} (\boldsymbol{X}^\top \boldsymbol{X}) (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I})^{-1}
\end{aligned}$$

Régression Ridge

Propriétés

Propriété 7. *Les estimateurs* $\widehat{\boldsymbol{\beta}}_{ridge}$ *et* $\widehat{\boldsymbol{\beta}}_{OLS}$ *vérifient :*

$$Cov\{\widehat{\boldsymbol{\beta}}_{OLS}\} - Cov\{\widehat{\boldsymbol{\beta}}_{ridge}\}$$
$$= \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I}\right)^{-1} \left[2n\lambda \boldsymbol{I} + n^2\lambda^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right] \left[\left(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I}\right)^{-1}\right]^\top$$

La matrice ci-dessus est définie positive, quel que soit $\lambda > 0$. Ceci signifie :

$$\mathrm{Cov}\{\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\} \succeq \mathrm{Cov}\{\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}\}$$

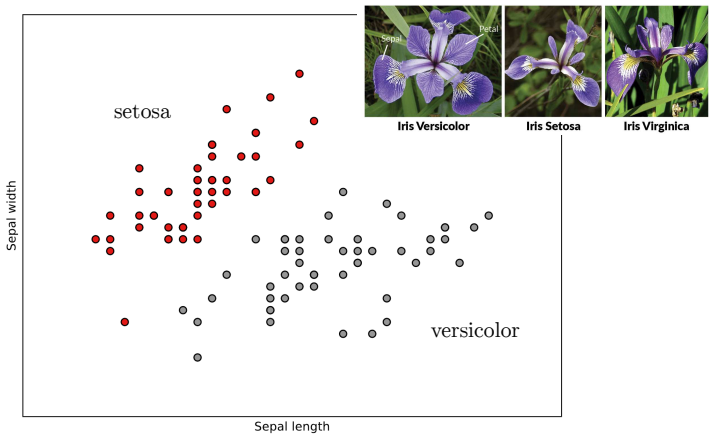
l'inégalité étant stricte pour $\lambda > 0$.

Régression logistique

Contexte

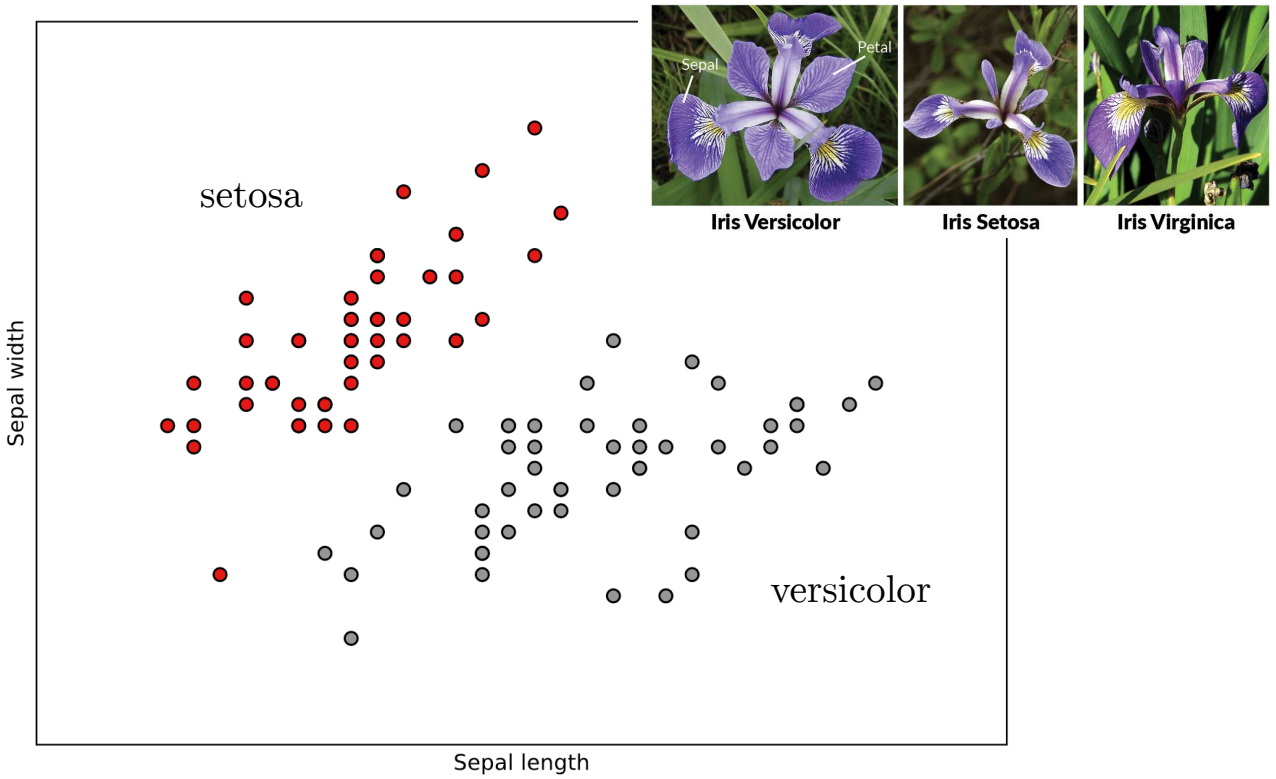
Problème : On s'intéresse à la classification automatique d'iris.
On dispose d'une base d'individus *Setosa* et *Versicolor*, caractérisés par les dimensions de leurs sépales et pétales.

Méthode : Le but de la régression logistique est d'effectuer une régression des individus $\boldsymbol{x}_i$ sur les étiquettes correspondantes $y_i \in \{-1, +1\}$.



Régression logistique

Contexte



Régression logistique

Spécification du modèle

Soit $y \in \{-1, +1\}$ la var. à prédire, et $\boldsymbol{x} \in \mathbb{R}^p$ le vecteur des var. explicatives. [a]

Notations :

- ▷ $P(y = \pm 1)$: probabilités a priori des classes, notées $P(\pm 1)$
- ▷ $P(\boldsymbol{x}|y = \pm 1)$: distrib. conditionnelles des observations, notées $P(\boldsymbol{x} \,|\, \pm 1)$

Principe :

La régression logistique repose sur l'hypothèse fondamentale que :

$$\ln \frac{P(\boldsymbol{x} \,|\, +1)}{P(\boldsymbol{x} \,|\, -1)} = a_0 + a_1 x_1 + \cdots + a_{p-1} x_{p-1}$$

Contrairement à l'AFD, on ne s'intéresse pas ici aux distributions conditionnelles mais à leur rapport.

a. Les variables y et $\boldsymbol{x}$ du problème sont aléatoires. Contrairement à l'usage, elles ne seront pas représentées par des lettres capitales $(Y, \boldsymbol{X})$ pour éviter toute confusion avec ce qui précède.

Régression logistique

Spécification du modèle

La spécification précédente peut être réécrite de manière équivalente :

$$\ln \frac{P(+1|\,\boldsymbol{x})}{1-P(+1|\,\boldsymbol{x})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$
$$= \boldsymbol{\beta}^\top \boldsymbol{x}$$

On désigne par *logit* de $P(+1\,|\boldsymbol{x})$ l'expression ci-dessus.

Il s'agit d'une régression logistique car le modèle ci-dessus provient d'une loi logistique. En effet, par manipulation du modèle ci-dessus, on obtient :

$$P(+1|\,\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{\beta}^\top \boldsymbol{x}}}$$

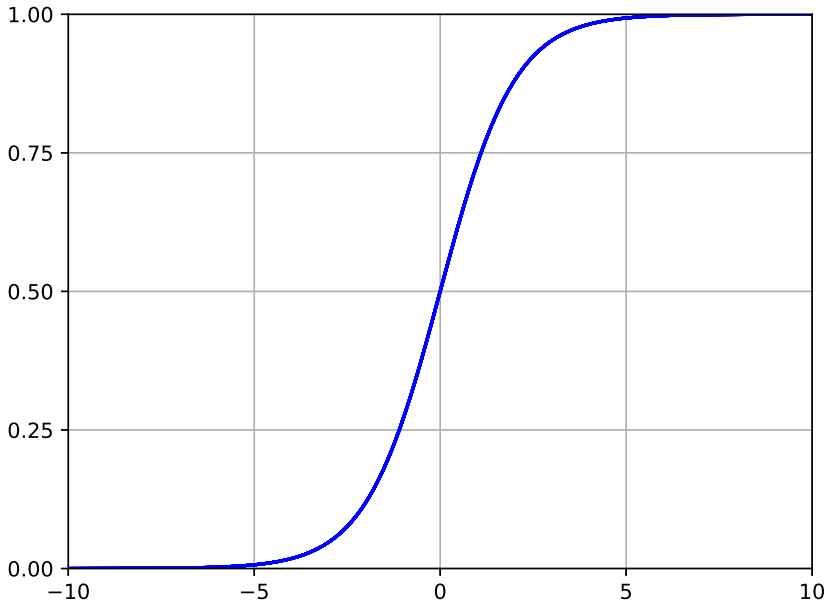
ainsi que :

$$P(-1|\,\boldsymbol{x}) = 1 - P(+1|\,\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}$$

Régression logistique

Fonction logistique

La loi logistique est définie par sa fonction caractéristique de même nom, appelée également sigmoïde.



Régression logistique

Règle de décision

Etant donné $\boldsymbol{\beta}$ estimé comme ci-après, la règle de décision mise en œuvre est :

- ▷ choisir $(+1)$ si $P(+1|\,\boldsymbol{x}) > P(-1|\,\boldsymbol{x})$
- ▷ choisir (-1) sinon

Régression logistique

Estimation des paramètres

Loi de y :

La probabilité d'appartenance d'un individu $\boldsymbol{x}$ à une classe $y = \pm 1$ est régit par une loi de Bernoulli :

$$P(y|\,\boldsymbol{x}) = P(+1|\,\boldsymbol{x})^{\frac{1+y}{2}} \times P(-1|\,\boldsymbol{x})^{\frac{1-y}{2}}, \qquad y \in \{-1, +1\}$$

Estimation de $\boldsymbol{\beta}$ par maximum de vraisemblance :

Les variables $\{y_i\}_{i=1}^n$ sont supposées i.i.d. La vraisemblance de $\boldsymbol{\beta}$ est donnée par :

$$L(y_1, \ldots, y_n \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i \,|\, \boldsymbol{x}_i)$$

qu'il faut maximiser par rapport à $\boldsymbol{\beta}$.

Régression logistique

Estimation des paramètres

Estimation de $\boldsymbol{\beta}$ par maximum de vraisemblance :

Afin de simplifier les calculs, on considère l'opposé de la log-vraisemblance, qu'il faut minimiser par rapport à $\boldsymbol{\beta}$:

$$
\begin{aligned}
-\ln L(y_1, \ldots, y_n \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\beta}) &= -\sum_{i=1}^{n} \ln P(y_i \,|\, \boldsymbol{x}_i) \\
&= \sum_{i=1}^{n} \ln \left[1 + e^{-y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i}\right] = J(\boldsymbol{\beta})
\end{aligned}
$$

La fonction coût $J(\boldsymbol{\beta})$ est différentiable sur $\mathbb{R}^d$ et strictement convexe. Elle admet un minimum global $\widehat{\boldsymbol{\beta}}$ satisfaisant la condition :

$$\nabla J(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$$

Régression logistique

Estimation des paramètres

$$\min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \triangleq \sum_{i=1}^{n} \ln\left[1 + e^{-y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i}\right]$$

Calcul du gradient :

La dérivée partielle de $J(\boldsymbol{\beta})$ par rapport à β_j est donnée par :

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{y_i \left[\boldsymbol{x}_i\right]_j p_i(\boldsymbol{\beta})}{1 + p_i(\boldsymbol{\beta})} \quad \text{avec} \quad p_i(\boldsymbol{\beta}) = e^{-y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i}$$

Ceci nous permet de réécrire le gradient de $J(\boldsymbol{\beta})$ sous la forme :

$$\nabla J(\boldsymbol{\beta}) = -\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}) \boldsymbol{y}$$

où $\boldsymbol{D}(\boldsymbol{\beta})$ est la matrice diagonale de termes diagonaux $[\boldsymbol{D}(\boldsymbol{\beta})]_{ii} = \frac{p_i(\boldsymbol{\beta})}{1+p_i(\boldsymbol{\beta})}$

Régression logistique

Estimation numérique des paramètres

La condition d'optimalité $\nabla J(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ n'admet pas de solution analytique. Il est nécessaire de recourir à une méthode numérique

Algorithme du gradient :
Choisir un point initial $\boldsymbol{\beta}_0$, un seuil ϵ, et un pas μ
Itérer les étapes suivantes à partir de $k = 0$

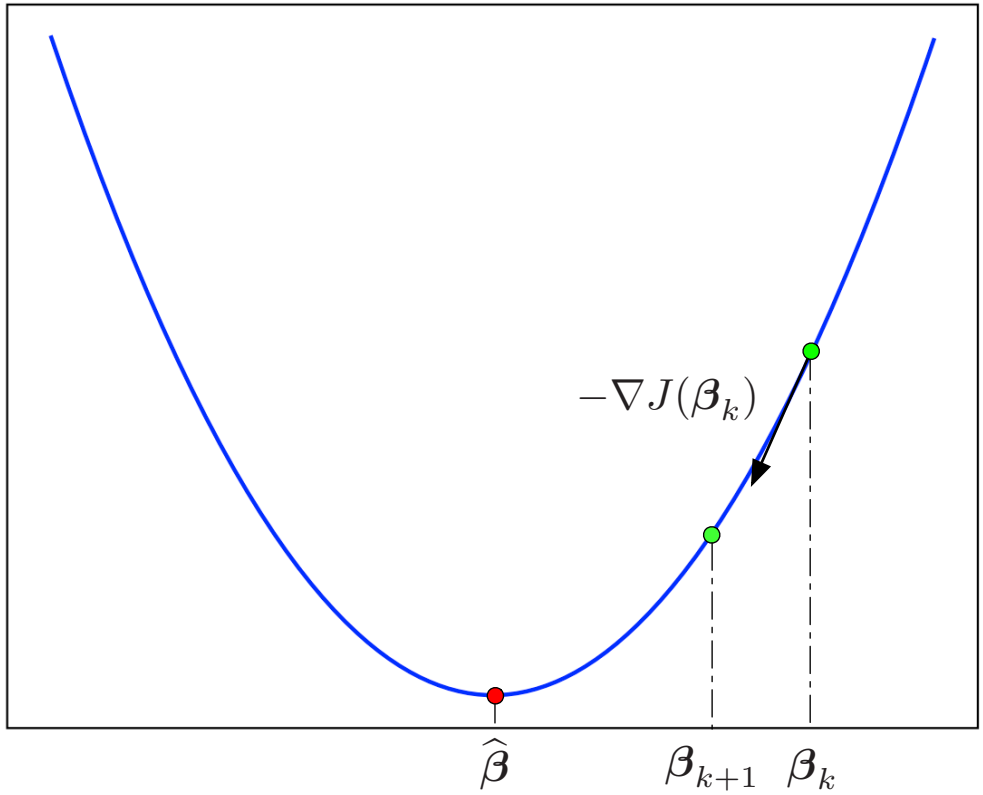
1. Calculer $\nabla J(\boldsymbol{\beta}_k)$
2. Test d'arrêt : si $\|\nabla J(\boldsymbol{\beta}_k)\| < \epsilon$, arrêt
3. Nouvel itéré : $\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \mu \nabla J(\boldsymbol{\beta}_k)$

Régression logistique :

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \mu\, \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k)\boldsymbol{y}$$

Régression logistique

Estimation numérique des paramètres



Régression logistique

Données iris