Machine Learning avancé
Modèles linéaires

Maxime Jumelle

ESLSCA Big Data - MBA 2

2019 - 2020

Sommaire

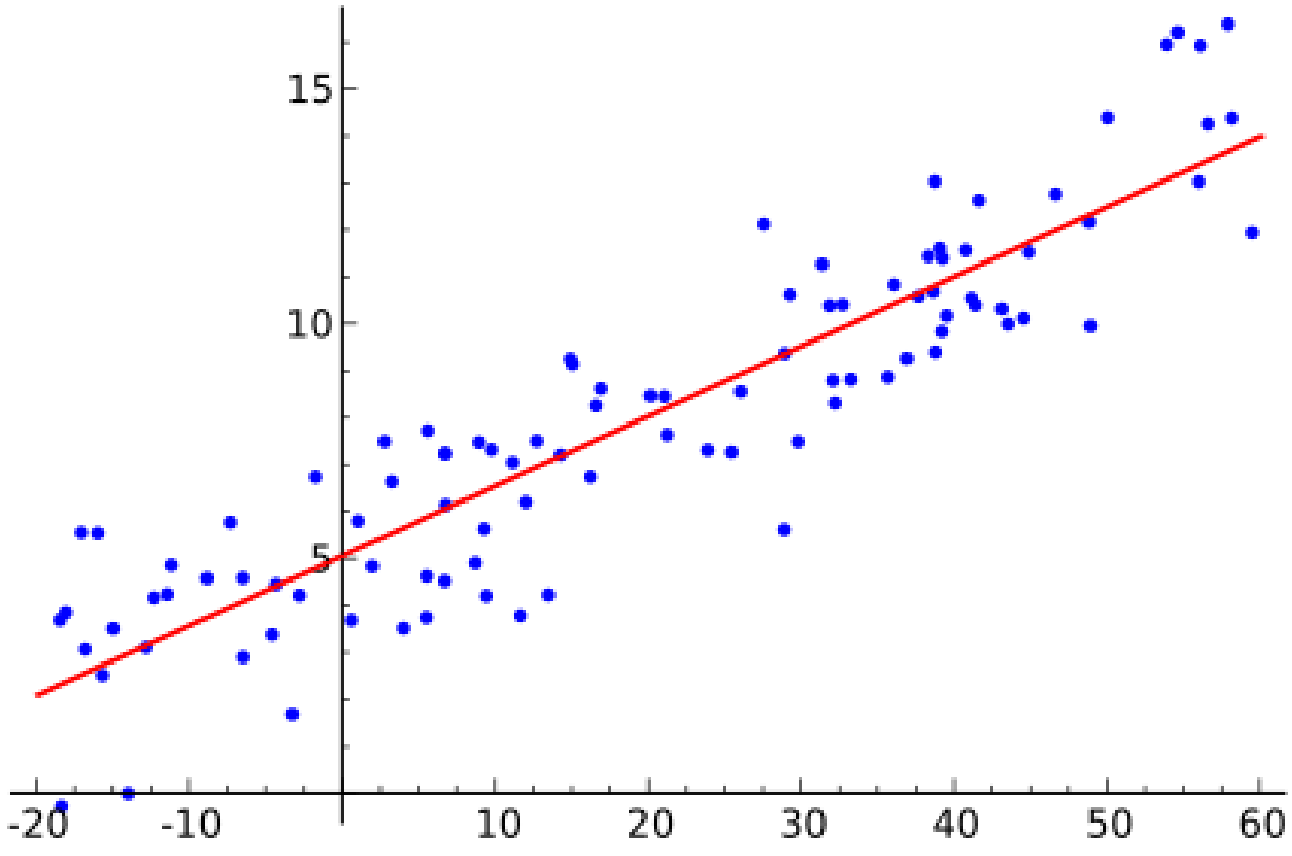
Fonctions affines

Considérons la cas classique d'une fonction affine :

$$y = ax + b$$

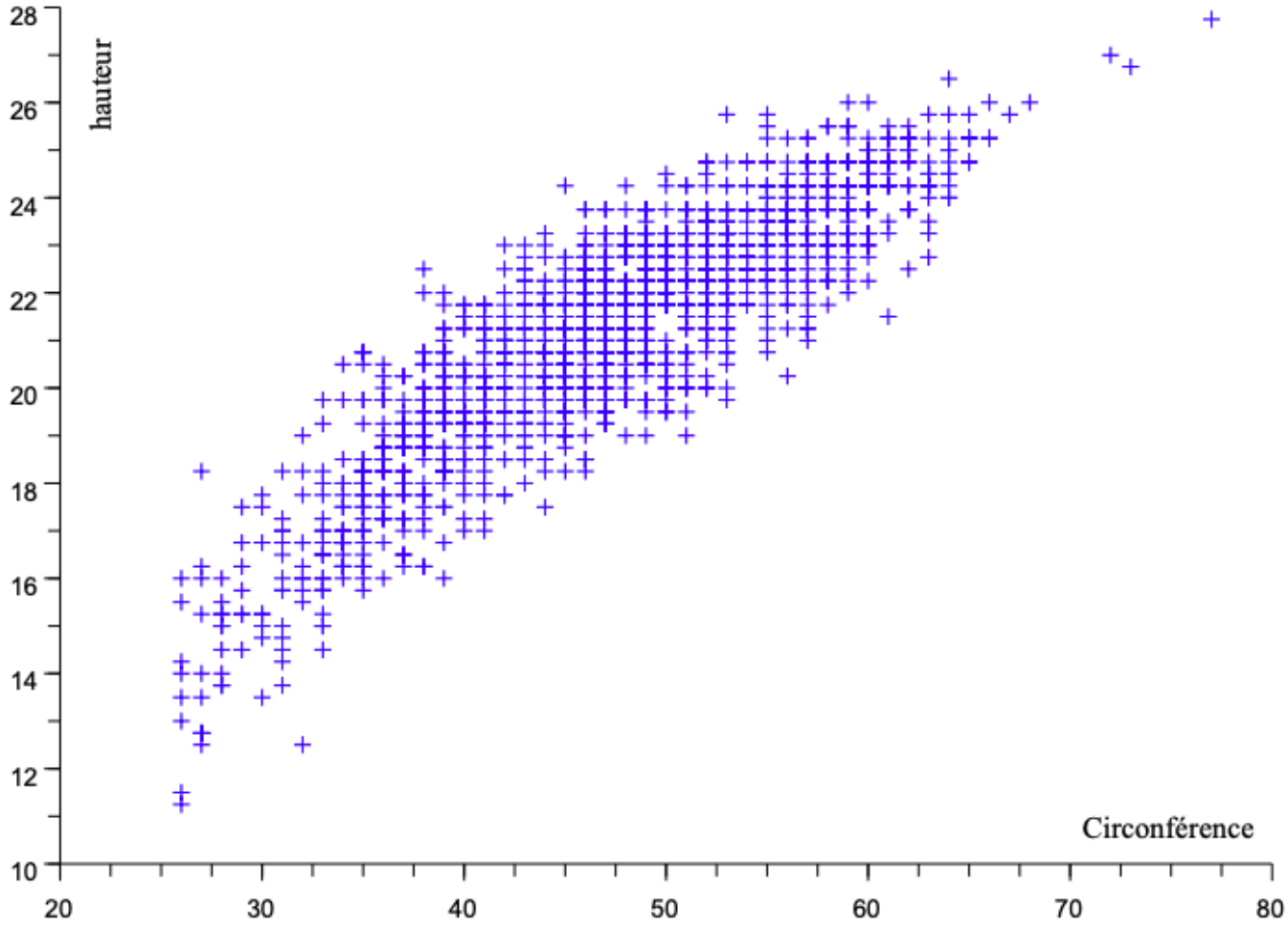
Ici, a et b sont des réels. Ces deux nombres définissent entièrement la courbe et permet donc d'obtenir une relation **affine** entre x et y. En statistique, cette relation est à la base des modèles dits **linéaires**, où une variable réponse se définit comme une somme de variables explicatives où chacune de ces dernières sont multipliés par un coefficient.

Fonctions affines



Exemple

Hauteur y (en m) d'eucalyptus en fonction de leur circonférence x (en m).



Modèle linéaire simple

Dans le modèle linéaire simple (une seule variable explicative), on suppose que la variable réponse suit le modèle suivant :

$$y_i = \beta_0 + \beta_1 \mathbf{x}_i + \varepsilon_i$$

On remarque la ressemblance avec la fonction affine présentée ci-dessus. La différence réside dans l'existence du terme aléatoire (appelé bruit) ε_i. Afin de considérer le modèle, il est nécessaire de se placer sous les hypothèses suivantes.

$$(\mathcal{H}) : \begin{cases} \mathbb{E}[\varepsilon_i] = 0 \\ \mathsf{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2 \end{cases}$$

La classe de fonctions candidates est

$$\mathcal{F} = \left\{ \mathbf{x} \mapsto \beta_0 + \beta_1 \mathbf{x} : (\beta_0, \beta_1) \in \mathbb{R}^2 \right\}$$

Modèle linéaire simple

Les différents éléments qui interviennent sont :

- β_0 : l'ordonnée à l'origine (nommée *intercept*)
- β_1 : le coefficient directeur
- $\mathbf{x}_i$: l'observation i
- y_i : le i-ème label observé
- ε_i : le bruit aléatoire lié à la i-ème observation

Pour un problème de régression, une bonne fonction de perte est la perte quadratique. Puisque nous sommes qu'à une seule dimension, cela revient à calculer le carré de la différence entre la label observé et la prédiction

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

dans l'optique d'estimer le modèle

$$\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

Modèle linéaire simple

Ainsi, la fonction objectif est naturellement défini comme la somme des erreurs quadratiques :

$$L(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Analytiquement, cela revient à résoudre le problème d'optimisation suivant

$$\min_{(\beta_0, \beta_1) \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \hat{f}(\mathbf{x}_i))^2 = \min_{(\beta_0, \beta_1) \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \mathbf{x}_i)^2$$

La solution peut se calculer facilement via les formules fermées suivantes :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Exercice

Démonter ce résultat. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$ dans le cas où $\bar{x} = 0$ et $\bar{y} = 0$.

Estimateurs

Attention

$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des **estimateurs** de β_0 et β_1 : on ne pourra jamais connaître la valeur des coefficients, on pourra seulement fournir des estimateurs !

- β_0 et β_1 : variables aléatoires
- $\hat{\beta}_0$ et $\hat{\beta}_1$: paramètres estimées du modèle

On défini ainsi les labels prédits $\hat{y}_1, \ldots, \hat{y}_n$ par

$$\hat{f}(\mathbf{x}_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_i$$

Bruits et résidus

Comment comparer les labels observés et prédits ? Dans le modèle, nous avons vu la présence de bruit aléatoire ε_i. Néanmoins, pour prédire une observation, ce bruit n'intervient pas. Comment pouvons-nous ainsi estimer ce bruit ?

$$\text{Résidus : } \hat{\varepsilon}_i = y_i - \hat{y}_i$$

Ces résidus permettent d'estimer la variance des bruits, et donc de pouvoir les caractériser :

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

Objectifs

Le modèle linéaire simple doit apprendre sur les données afin de déterminer des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ en minimisant la fonction de perte. Très souvent en régression, on minimise la perte quadratique moyenne (RMSE en anglais).

$$RMSE(Y, \hat{Y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \frac{1}{\sqrt{n}}\|Y - \hat{Y}\|_2$$

La RMSE mesure l'écart global entre les prédictions et les observations. Plus la RMSE est proche de 0, mieux le modèle prédit correctement les données conformément aux observations.

Exercice

Montrer que

$$RMSE^2(Y, \hat{Y}) = \frac{\|\hat{\varepsilon}\|^2}{n}$$

Score

Pour le modèle linéaire simple, une étude visuelle nous permet d'établir plus ou moins la performance du modèle.

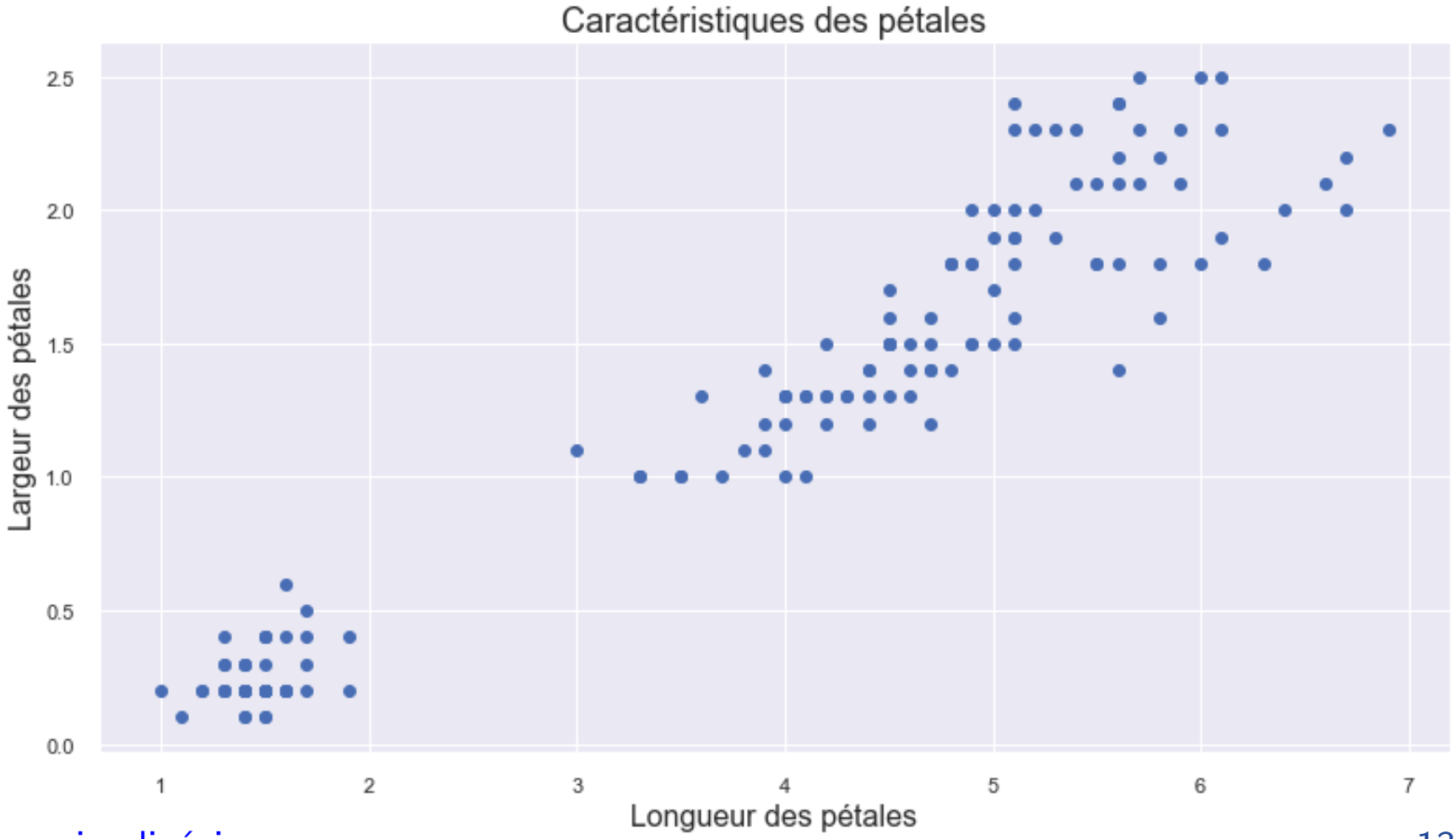
En utilisant la définition des résidus, on peut définir le coefficient de détermination $R^2 \in [0, 1]$.

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

- R^2 proche de 1 : fort pouvoir explicatif
- R^2 proche de 0 : faible pouvoir explicatif

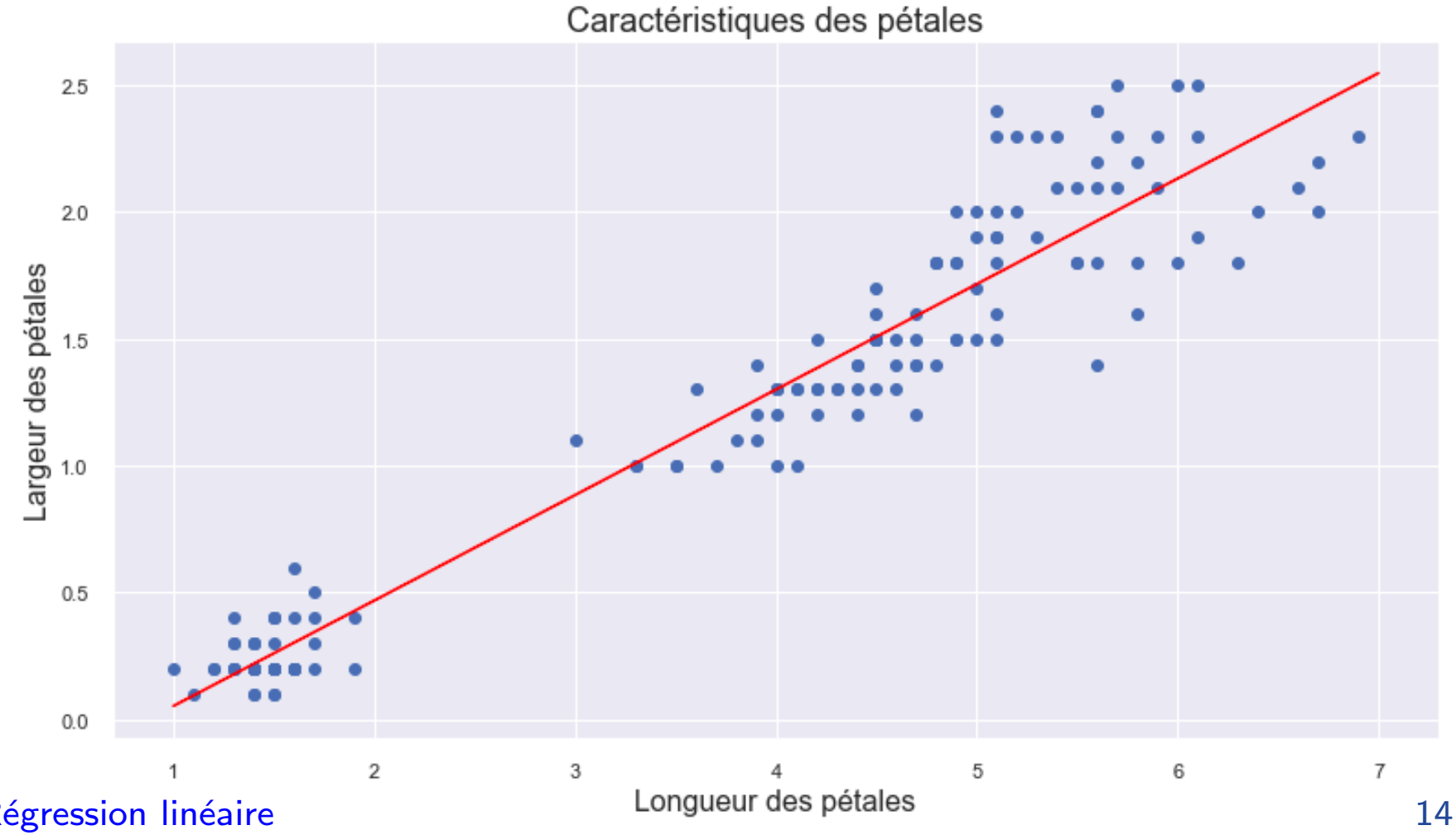
Exemple

Considérons des fleurs d'iris. On souhaite modéliser une relation linéaire entre les longueur des pétales et les largeurs des pétales.



Exemple

$$R^2 = 92.7\%$$

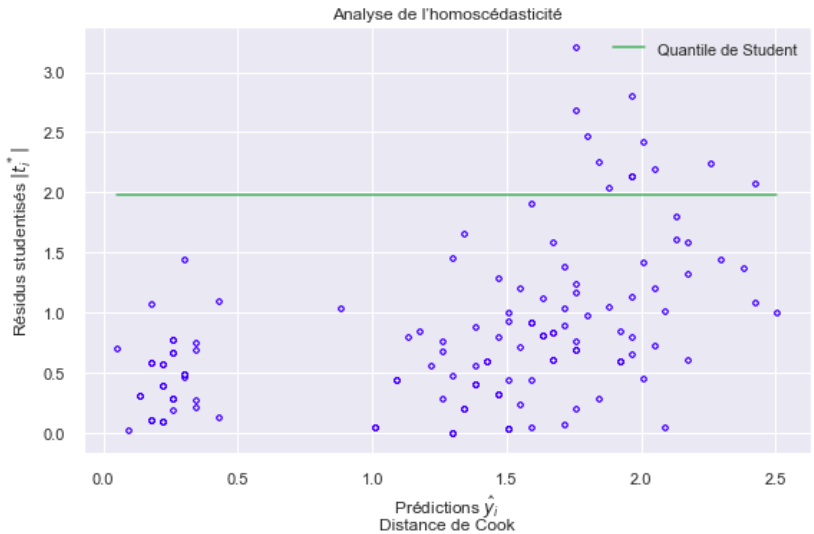


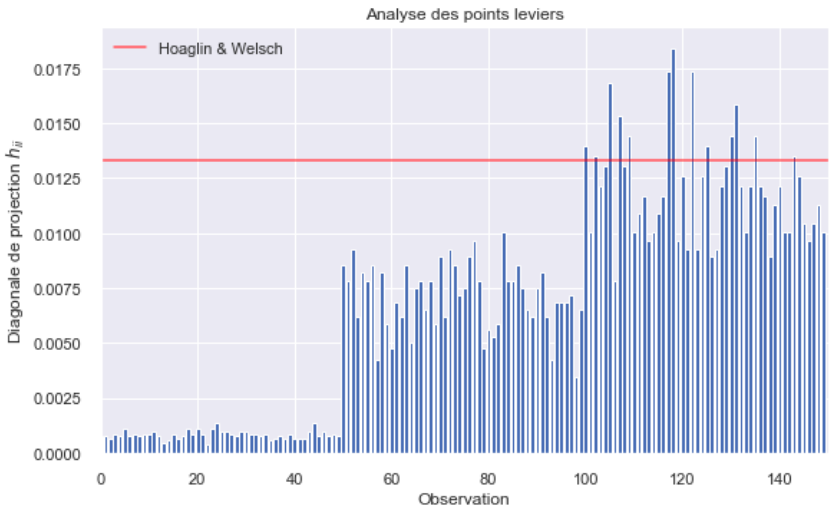
Exemple

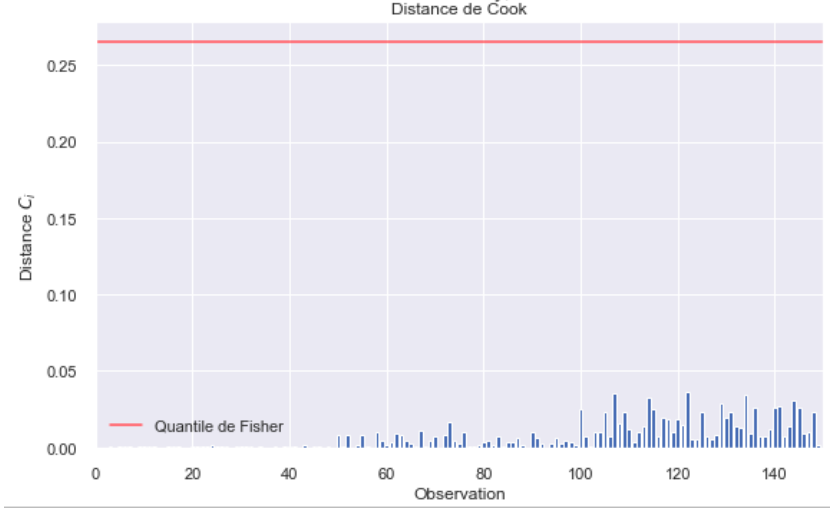
La relation linéaire obtenue par le modèle empirique est :

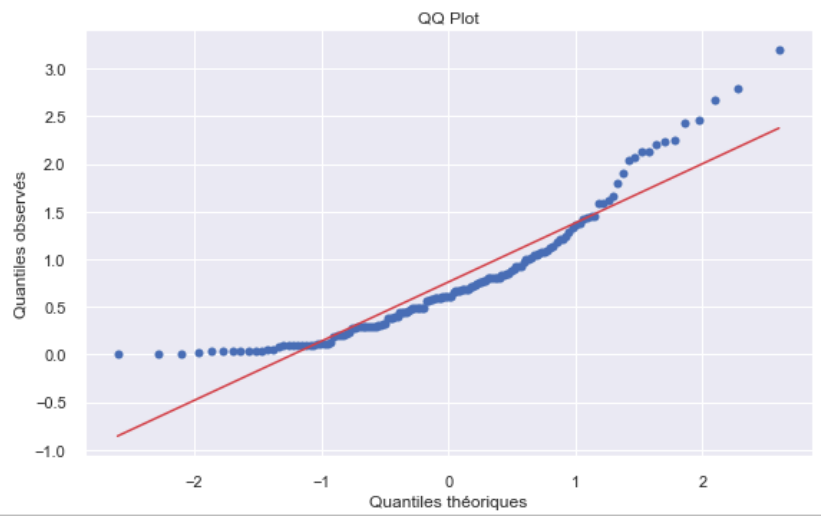
$$y = -0.36 + 0.42x$$

Exemple


Analyse de l'homoscédasticité
Quantile de Student
Résidus studentisés $|t_i^*|$
Prédictions $\hat{y}_i$


Analyse des points leviers
Hoaglin & Welsch
Diagonale de projection h_{ii}
Observation


Distance de Cook
Quantile de Fisher
Distance C_i
Observation


QQ Plot
Quantiles observés
Quantiles théoriques

Modèle linéaire multiple

Dans le cas multiple (pour p variables explicatives), pour la i-ème observation, le modèle s'écrit :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_{ij} + \varepsilon_i$$

Ainsi, une observation $\mathbf{x}_i$ n'est plus une valeur, mais un **vecteur** $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$. Il est plus commode de regrouper les labels observés y_i et ces vecteurs d'observations $\mathbf{x}_i$ dans des matrices :

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

Sous les hypothèses équivalentes du modèle simple en plus grande dimension

$$(\mathcal{H}) : \begin{cases} \text{rank}(X) = p \\ \mathbb{E}[\varepsilon] = 0 \text{ et } \text{Var}(\varepsilon) = \sigma^2 I_p \end{cases}$$

Modèle linéaire multiple

La classe de fonctions candidates est

$$\mathcal{F} = \left\{ \mathbf{x} \mapsto \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_{\bullet j} : (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1} \right\}$$

Les différents éléments qui interviennent sont :

- β : le vecteur directeur
- X : la matrice des observations
- Y : le vecteur des labels observés
- ε : le vecteur de bruit

Avec $X = (\mathbf{1}, X_1, \ldots, X_n)$, $Y = (y_1, \ldots, y_n)^\top$ et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. La solution des MCO (Moindres Carrés Ordinaires) est alors :

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

Modèle linéaire multiple

Preuve

Soit la fonction de coût à minimiser s'écrit

$$S(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2$$

S étant quadratique (donc convexe et différentiable), la solution existe et est unique.

$$\nabla S(\beta) = -X^\top (Y - X\beta) = -X^\top Y + X^\top X\beta$$

La solution $\hat{\beta}$ annule le gradient de la fonction de coût, ce qui nous amène à résoudre une équation linéaire.

$$\hat{\beta} = \text{argmin } \nabla S(\beta) \Leftrightarrow X^\top X\hat{\beta} = X^\top Y$$

Or la matrice $X^\top X$ est inversible sous l'hypothèse $(\mathcal{H})$ d'où la formule fermée

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

Bruits et résidus

Les résidus s'obtiennent par le même calcul que pour la régression linéaire simple.

$$\text{Résidus : } \hat{\varepsilon}_i = y_i - \hat{y}_i$$

En revanche, l'estimateur de la variance dépends cette fois-ci du nombre de variables p :

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

Score

Pour généraliser la notion de coefficient de détermination R^2 à une plus grande dimension, on introduit les quantités suivantes :

- $SCT = \|Y - \bar{y}\mathbf{1}\|^2$ est la somme des carrés totaux
- $SCE = \|\hat{Y} - \bar{y}\mathbf{1}\|^2$ est la somme des carrés expliqués
- $SCR = \|\hat{\varepsilon}\|^2$ est la somme des carrés résiduels

$$R^2 = 1 - \frac{SCR}{SCT} = \frac{SCE}{SCT}$$

Exercice

Retrouver la formule du R^2 dans le cas de la régression simple à partir de la formule ci-dessus.

Interprétation géométrique

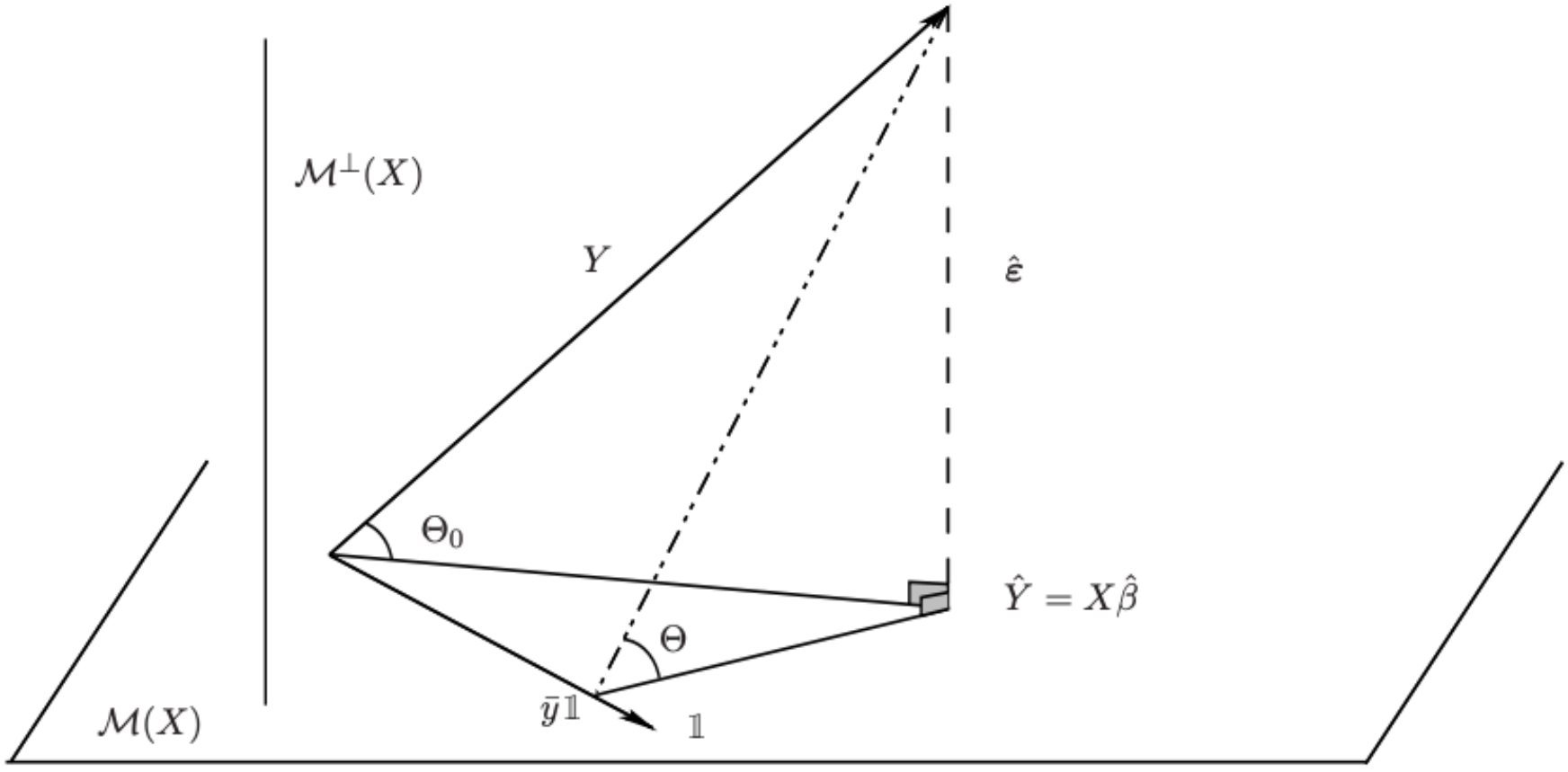


FIGURE – Source : Régression linéaire - Arnaud Guyader

Régression pénalisée

Lorsque la dimension (le nombre de variables) devient important par rapport aux données, l'hypothèse fondamentale $n > p$ n'est plus vérifiée. Dans ce cas, il existe plusieurs approches pour effectuer une régression linéaire sur ces données :

- Effectuer une réduction de dimension sur les données : en pratique, d'importantes conditions sur les données sont requises ou bien les temps de calculs sont trop importants.
- Sélectionner au préalable un sous-ensemble des variables.

Cette sélection de variable peut se faire par l'utilisateur, ou alors "automatiquement" en agissant directement sur la fonction de perte : c'est la **régularisation**.

Ridge

L'estimateur Ridge $\hat{\beta}^{\text{Ridge}}$ est la quantité qui minimise la fonction suivante :

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

Avec $\lambda \geq 0$.

Il existe une solution au problème précédent : l'estimateur Ridge existe et vaut

$$\hat{\beta}^{\text{Ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y$$

LASSO

L'estimateur LASSO $\hat{\beta}^{\text{LASSO}}$ est la quantité qui minimise la fonction suivante :

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Avec $\lambda \geq 0$.

Attention

La fonction à minimiser n'est plus différentiable !

LASSO

Exercice

Montrer l'équivalence des deux problèmes d'optimisation :

$$(P1) \quad \hat{\beta} \in \min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$(P2) \quad \hat{\beta} \in \min_{\|\beta\|_1 \leq t} \frac{1}{2}\|Y - X\beta\|_2^2$$

Pour un certain t dépendant de λ (indice : utiliser les conditions de Karush-Kuhn-Tucker).

LASSO

L'estimateur LASSO est difficile à calculer par rapport à l'estimateur Ridge. De plus, du fait de la non différentiabilité de la fonction de perte, l'unicité de la solution n'existe que sous certaines conditions.

On montre que l'algorithme de descente du gradient n'est pas optimal dans ce cas. On utilise alors plus souvent **Coordinate Descent** (implémenté sous *scikit-learn*).

Une solution exacte

Supposons que les colonnes de X soient orthonormales ($\langle \mathbf{x}_i, X_j \rangle = \delta_{ij}$). Alors on peut trouver une formule fermée pour $\hat{\beta}^{\mathsf{LASSO}}$ à partir de la solution des MCO $\hat{\beta}$:

$$\hat{\beta}_j^{\mathsf{LASSO}} = \mathsf{sgn}(\hat{\beta}_j) \max(|\hat{\beta}_j| - \lambda, 0)$$

Sommaire

Définition

La régression logistique (ou *modèle binomial*) est un cas particulier de modèle linéaire généralisé (GLM).

Bien qu'il s'agisse d'une régression, la tâche effectuée est en réalité une classification.

Formalisme

On dispose de n observations à p variables continues ou binaires, et l'objectif est de prédire la variable réponse $Y \in \{0, 1\}$. On note $X = (X_1, \ldots, X_p)$ ces variables, ainsi que $x_j, 1 \leq j \leq p$:

- à valeurs dans $\mathbb{R}$ si X_j est continue.
- à valeurs dans $\{0, 1\}$ si X_j est binaire.

Notation

Comme convenu, on utilisera

- x_j pour désigner la valeur de la j-ème d'un individu quelconque
- $\mathbf{x}_i$ le vecteur de taille $p + 1$ associé au i-ème individu **avec un premier coefficient égal à** 1.

Formalisme

La similitude avec le modèle linéaire se situe dans la relation entre la probabilité et les variables. Néanmoins, là où pour la régression linéaire, la variable réponse Y était à valeurs dans $\mathbb{R}$, le modèle ici est différent étant donné que les probabilités sont bornées entre 0 et 1. Il n'est donc pas possible de lier directement la probabilité au variables par une relation linéaire.

Afin de *mapper* la réponse sur l'intervalle $[0, 1]$, on utilise la fonction logit, qui possède les bonnes propriétés de bijectivité et dérivabilité.

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

Formalisme

On souhaite modéliser la distribution *a posteriori* $Y|X = \mathbf{x}$ d'une observation $\mathbf{x}$. Notons que la loi Y étant binaire, connaître l'événement $\{Y = 1|X = \mathbf{x}\}$ revient à connaître entièrement la distribution de la loi a posteriori.

Le modèle **logit** établit la relation suivante entre la loi a posteriori et une combinaison linéaire des observations.

$$\log \frac{\mathbb{P}(Y = 1|X = \mathbf{x})}{1 - \mathbb{P}(Y = 1|X = \mathbf{x})} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \beta^\top \mathbf{x}$$

Exercice

Montrer que

$$\mathbb{P}(Y = 1|X = \mathbf{x}) = \frac{\exp(\beta^\top \mathbf{x})}{1 + \exp(\beta^\top \mathbf{x})}$$

Estimation

Attention

Tout comme pour la régression multiple, le coefficient *intercept* β_0 est implicitement inclu dans l'observation $\mathbf{x}$, mais qui n'est à considérer **uniquement dans le cas de l'estimation** !

On cherche à estimer les coefficients β. La solution exacte consiste à utiliser l'estimateur du maximum de vraisemblance :

$$L(X; Y, \beta) = \prod_{i=1}^{n} \mathbb{P}(Y = 1|X = \mathbf{x}_i)^{y_i} \mathbb{P}(Y = 0|X = \mathbf{x}_i)^{1-y_i}$$

Souvent, on préfère minimiser la log-vraisemblance, car plus simple numériquement :

$$\log L(X; Y, \beta) = \sum_{i=1}^{n} y_i \log \mathbb{P}(Y = 1|X = \mathbf{x}_i) + (1-y_i) \log \mathbb{P}(Y = 0|X = \mathbf{x}_i)$$

Sauf que ... trop complexe à calculer. Les algorithmes implémentés calculent une solution approchée, notamment avec Newton-Raphson.

Descente de gradient appliquée

On rappelle que

$$L(X;Y,\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta^\top \mathbf{x}_i)}{1+\exp(\beta^\top \mathbf{x}_i)} \right)^{y_i} \left(\frac{1}{1+\exp(\beta^\top \mathbf{x}_i)} \right)^{1-y_i}$$

Exercice

1. Montrer que

$$\log L(X;Y,\beta) = \sum_{i=1}^{n} \left\{ y_i \beta^\top \mathbf{x}_i - \log(1+\exp(\beta^\top \mathbf{x}_i)) \right\}$$

2. Calculer le gradient $\nabla_\beta \log L(X;Y,\beta)$ de la log-vraisemblance.
3. En déduire un algorithme de descente de gradient pour la régression logistique.

Prédiction

Une fois les β estimés, comment prédire une nouvelle observation $\mathbf{x}$?
Rappelons que nous disposons de la probabilité suivante :

$$\mathbb{P}(Y = 1|X = \mathbf{x}) = 1 - \mathbb{P}(Y = 0|X = \mathbf{x})$$

Grâce à la règle de Bayes, on peut construire le classifieur $\hat{g}$ suivant :

$$\hat{g}(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = \mathbf{x}) > \mathbb{P}(Y = 0|X = \mathbf{x}) \\ 0 & \text{sinon} \end{cases}$$

Par calcul de la loi a posteriori, on obtient le classifieur suivant.

$$\hat{g}(\mathbf{x}) = \begin{cases} 1 & \exp(\beta^\top \mathbf{x}) > 1 \\ 0 & \text{sinon} \end{cases}$$

Exercice

Déterminer la frontière du classifieur :

$$\{\mathbf{x} \in \mathbb{R}^{p+1} : \exp(\beta^\top \mathbf{x}) = 1\}$$

Interprétation

Contraitement au modèle linéaire, il ne suffit pas de multiplier les valeurs par les coefficients pour obtenir une interprétation : en effet, la linéarité est effective **par transformation logistique du rapport des probabilités**.
Il est donc nécessaire de calculer convenablement ces coefficients en prenant en compte la transformation logistique que l'on a appliqué.
Rappelons le modèle :

$$\log \frac{\mathbb{P}(Y=1|X=\mathbf{x})}{1-\mathbb{P}(Y=1|X=\mathbf{x})} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \beta^\top \mathbf{x}$$

Un simple passage par l'exponentielle nous donne

$$\text{ratio} = \frac{\mathbb{P}(Y=1|X=\mathbf{x})}{1-\mathbb{P}(Y=1|X=\mathbf{x})} = \exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right\}$$

Interprétation

Que se passe-t-il au niveau du ratio si l'on ajoute une unité sur la variable j ?

$$\frac{\text{ratio}_{x_j+1}}{\text{ratio}} = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \beta_j\right\}}{\exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right\}} = \exp(\beta_j)$$

En résumé, chaque variable agit exponentiellement sur le rapport des probabilités : une augmentation d'une unité sur la variable x_j augmente exponentiellement le rapport des probabilités en le coefficient associé. Par exemple, si $\beta_j = 0.8$, alors une augmentation d'une unité entraîne le rapport suivant :

$$\text{ratio}_{x_j+1} = 2.23 \times \text{ratio}$$

Cas pratique

Réalisons une régression logistique sur un jeu de données. Supposons que l'on souhaite évaluer la qualité d'un modèle de voiture en fonction de ses caractéristiques. On dispose de deux classes : 0 pour une qualité basse à moyenne, et 1 pour une qualité de bonne à excellente. Les informations sur le jeu de données sont :

- 1727 modèles de voitures.
- 6 variables explicatives, dont 2 quantitatives et 4 qualitatives.
- On dispose de 1209 voitures de qualité basse/moyenne et 518 voitures de qualité bonne/excellente.

Variables explicatives

Les variables explicatives sont :

- **buying** (qualitative) : un prix d'achat (bas, moyen, haut et très haut).
- **maint** (qualitative) : entretien nécessaire pour le fonctionnement (bas, moyen, haut et très haut).
- **doors** (quantitative) : le nombre de portes.
- **persons** (quantitative) : le nombre de personnes.
- **lug_boot** (qualitative) : la taille du coffre (petit, moyen et grand).
- **safety** (qualitative) : le niveau de sécurité (faible, modéré et fort).

La variable que l'on souhaite modéliser (i.e. la variable **réponse**) est nommée **eval**.

Tableau de données

	buying	maint	doors	persons	lug_boot	safety	eval
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc
5	vhigh	vhigh	2	2	big	low	unacc
6	vhigh	vhigh	2	2	big	med	unacc
7	vhigh	vhigh	2	2	big	high	unacc

Données encodées

Dans la pratique, le modèle n'est pas capable d'effectuer des calculs avec des catégories : **il est nécessaire d'encoder les données**. Ici, on a fait correspondre, pour chaque modalité, un nombre associé qui permet de les distinguer lors de l'optimisation du modèle.

	buying	maint	doors	persons	lug_boot	safety	eval
0	3	3	2	2	0	1	0
1	3	3	2	2	0	2	0
2	3	3	2	2	1	0	0
3	3	3	2	2	1	1	0
4	3	3	2	2	1	2	0
5	3	3	2	2	2	0	0
6	3	3	2	2	2	1	0
7	3	3	2	2	2	2	0

Interprétation

Le F1 score obtenu est de 77%, ce qui est acceptable. Le coefficient *intercept* estimé vaut -8.24 et les coefficients $\hat{\beta}$ estimés sont :

$$\begin{pmatrix} -0.89 & -0.79 & 0.27 & 1.12 & 0.76 & 2.58 \end{pmatrix}$$

Précedemment, nous avons vu que c'était l'exponentielle des coefficients qui impactaient directement les ratios de probabilité :

$$\frac{\text{ratio}_{x_j+1}}{\text{ratio}} = \exp(\beta_j)$$

Pour pouvoir correctement interpréter, il faut donc composer par une exponentielle le vecteur :

$$\begin{pmatrix} 0.41 & 0.45 & 1.31 & 3.07 & 2.13 & 13.18 \end{pmatrix}$$

Interprétation

Rapports de ratio :

- 0.41 : buying
- 0.45 : maint
- 1.31 : doors
- 3.07 : persons
- 2.13 : lug_boot
- 13.18 : safety

Interprétation

Un des critères les plus discriminants est sans doute le **niveau de sécurité de la voiture**. Ensuite, le nombre de personnes est le second critère le plus important dans la prédiction de la qualité d'une voiture.

En revanche, le prix d'achat possède un effet inverse : il semblerait que le prix d'une voiture joue dans le sens inverse sur la qualité d'une voiture, de même que la maintenance/entretien nécessaire. Cela peut s'expliquer par des **biais humains** : on est beaucoup plus attentif et sélectif sur une voiture dont le prix de vente est élevé.

Interprétation

De même, il est important de s'interroger sur l'origine de la variable **safety**. En effet il est probable que cette dernière ait été crée par une personne intermédiaire entre le constructeur et l'acheteur (potentiellement un revendeur), ce qui peut fausser son interprétation car :

- On ne sait pas quelles ont été les règles permettant de choisir le niveau de sécurité.
- Un biais humain peut également exister lors de la création de ces règles et/ou lors de l'attribution des niveaux de sécurité.