

Régression linéaire et logistique

Pierre Gaillard

21 septembre 2017

Dans ce cours, nous allons étudier le problème simple mais encore abondamment utilisé de nos jours : la régression¹ linéaire. Nous finirons par un algorithme de l'état de l'art pour la classification : la régression logistique. Ce cours est basé sur les livres RIVOIRARD et STOLTZ, 2012 et CORNILLON et MATZNER-LØBER, 2011, ainsi que le cours des années précédentes de Simon Lacoste-Julien sur la régression logistique.

1 Régression linéaire

Commençons par un exemple de problème pratique. Pour mieux optimiser sa production, un producteur s'intéresse à modéliser la consommation électrique en France en fonction de la température (cf. Figure 1).

L'objectif est de trouver une fonction f telle qui explique bien la consommation électrique $(y_i)_{1 \leq i \leq n}$ en fonction de la température $(x_i)_{1 \leq i \leq n}$, soit $y_i \approx f(x_i)$. Pour cela, on peut choisir un espace de fonction \mathcal{F} et résoudre le problème de minimisation du risque empirique :

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{\mathcal{R}}_n(f) := \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2. \quad (1)$$

Il faut faire attention au choix de l'espace de fonction pour éviter le sur-apprentissage (cf. Figures 1). Bien que l'erreur quadratique moyenne diminue quand l'espace \mathcal{F} s'agrandit (degrés de polynômes plus grands), l'estimateur \hat{f} perd son pouvoir prédictif. L'espace des fonctions linéaires de la

1. Le mot « régression » aurait été introduit par Galton au XIXe siècle. En modélisant la taille des individus en fonction de celle de leurs pères, Galton a observé un retour (« regression » en anglais) vers la taille moyenne. Les pères plus grands que la moyenne ont tendance à avoir des enfants plus petits qu'eux et vice-versa pour les pères plus petits.

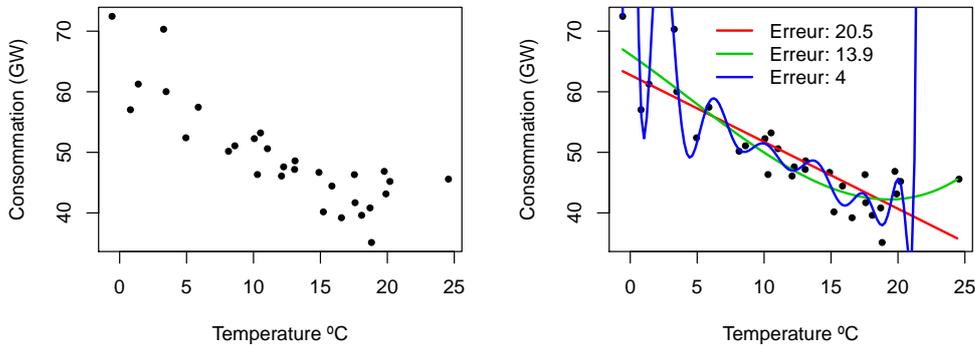


FIGURE 1 – Consommation électrique française (GW) en fonction de la température (°C). À droite, sont tracés fonctions minimisant l’erreur pour les espace de polynômes de degrés 1 (rouge), 3 (vert) et 30 (bleu)

forme $f : x \mapsto ax + b$ est le plus simple. C’est celui que nous allons étudier dans ce cours.

1.1 Le modèle linéaire

On dispose d’un vecteur d’observations $Y := (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ et de $p \geq 1$ vecteurs de variables dites explicatives $X_1 := (x_{1,1}, \dots, x_{n,1})^\top, \dots, X_p := (x_{1,p}, \dots, x_{n,p})^\top \in \mathbb{R}^n$. Un modèle de régression linéaire est défini par une équation de la forme :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i, \quad (2)$$

où $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$ est un vecteur d’erreurs (ou de bruit). Il vient du fait qu’en pratique les observations y_i ne collent jamais complètement à la prévision linéaire. On suppose qu’il s’agit d’un n -échantillon (i.e., les ε_i sont i.i.d.) d’espérance nulle $\mathbb{E}[\varepsilon_i] = 0$. Souvent, il est supposé de loi Gaussienne $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Pour simplifier les notations, nous supposons que le premier vecteur des variables explicatives est le vecteur constant $X_1 = (1, \dots, 1)^\top$ de sorte que le coefficient β_0 de l’équation (2) n’est plus nécessaire. Nous définissons la matrice des variable explicatives

$$X := \begin{bmatrix} 1 & x_{1,2} & \dots & x_{1,p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n,2} & \dots & x_{n,p} \end{bmatrix}$$

composée du vecteur colonne constant $X_1 = (1, \dots, 1)^\top$ et des colonnes des variables explicatives X_2, \dots, X_p . Nous pouvons réécrire la définition du modèle linéaire (2) de façon matricielle.

Definition 1.1 (Modèle linéaire). *Un modèle linéaire se définit par une équation de la forme :*

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

où $Y \in \mathbb{R}^n$ est le vecteur des observations, $X \in \mathbb{R}^{n \times p}$ est la matrice de variables explicatives et $\beta \in \mathbb{R}^p$ le vecteur de coefficients à estimer. On suppose que la matrice X est injective (i.e., $\text{rg}(X) = p$) et que le vecteur de bruit $\varepsilon \in \mathbb{R}^n$ est un n -échantillon centré ($\mathbb{E}[\varepsilon] = 0$ et $\text{Var}[\varepsilon] = \sigma^2 I_n$).

Remarque : dans toute la suite, bien que X puisse être aléatoire, il s'agit en pratique de variables explicatives observées (ou calculées) par le prévisionniste avant d'effectuer une prévision. On supposera donc qu'il s'agit de.

1.2 L'estimateur des moindres carrés ordinaire (OLS)

Les points (X_i, y_i) étant observés, l'objectif est de trouver la fonction affine $\hat{f} : x \mapsto x\hat{\beta}$ minimisant l'erreur quadratique moyenne (1).

Definition 1.2 (Estimateur des moindres carrés ordinaires (OLS)). *On appelle estimateur des moindres carrés ordinaire, le minimiseur $\hat{\beta} \in \mathbb{R}^{p+1}$ du risque empirique :*

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \beta_0 + \sum_{j=1}^p \beta_j x_{i,j} \right)^2 = \arg \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|. \quad (3)$$

où $\|\cdot\|$ est la norme euclidienne.

Proposition 1.1. *Sous les hypothèses et les notations de la Définition 1.1, l'OLS existe et est unique. Il s'écrit $\hat{\beta} = (X^\top X)^{-1} X^\top Y$.*

Démonstration. Par définition, l'OLS est le minimiseur du risque empirique $\hat{\mathcal{R}}_n : \mathbb{R}^p \rightarrow \mathbb{R}_+$ définie par : pour tout $\beta \in \mathbb{R}^p$

$$\hat{\mathcal{R}}_n(\beta) = \frac{1}{n} \|Y - X\beta\|_2^2 = \frac{1}{n} (\beta^\top (X^\top X) \beta - 2\beta^\top X^\top Y + \|Y\|^2).$$

Pour être un minimum, il est nécessaire que le gradient s'annule. Celui-ci s'écrit :

$$\nabla \hat{\mathcal{R}}_n(\hat{\beta}) = \frac{1}{n} (\hat{\beta}^\top (X^\top X) + (X^\top X) \hat{\beta} - 2X^\top Y) = \frac{2}{n} ((X^\top X) \hat{\beta} - X^\top Y).$$

où la dernière égalité est parce que la matrice $X^\top X \in \mathbb{R}^{p \times p}$ est symétrique. Comme X est de rang p par hypothèse, $X^\top X$ est même symétrique définie

positive et inversible (laissé en exercice). On en déduit donc qu'un optimum de $\widehat{\mathcal{R}}_n$ vérifie nécessairement :

$$\widehat{\beta} = (X^\top X)^{-1}XY.$$

Il reste cependant à vérifier qu'il s'agit bien d'un minimum et donc que la Hessienne est définie positive. Ce qui est le cas car : $\nabla^2 \widehat{\mathcal{R}}_n(\widehat{\beta}) = \frac{2}{n}(X^\top X)$. \square

Une interprétation géométrique Le modèle linéaire cherche à modéliser le vecteur $Y \in \mathbb{R}^n$ des observations par une combinaison linéaire de la forme $X\beta \in \mathbb{R}^n$. L'image de X est l'espace des solutions, noté $\text{Im}(X) = \{Z \in \mathbb{R}^n : \exists \beta \in \mathbb{R}^p \text{ tq } Z = X\beta\} \subseteq \mathbb{R}^n$. C'est le sous-espace vectoriel de \mathbb{R}^n engendré par les $p < n$ colonnes de la matrice de variables explicatives. Comme $\text{rg}(X) = p$, il est de dimension p . Selon la définition du modèle linéaire, Y est la somme d'un élément de $\text{Im}(X)$ et d'un bruit. En minimisant $\|Y - X\beta\|$ (cf. Définition 1.2), on cherche donc l'élément de $\text{Im}(X)$ le plus proche de Y . Il s'agit du projeté orthogonal de Y sur $\text{Im}(X)$, noté \widehat{Y} . Par définition de l'OLS et par la Proposition 1.1, on a :

$$\widehat{Y} \stackrel{\text{Déf 1.2}}{=} X\widehat{\beta} \stackrel{\text{Prop. 1.1}}{=} X(X^\top X)^{-1}X^\top Y.$$

En particulier, $P_X := X(X^\top X)^{-1}X^\top$ est la matrice de projection orthogonal sur $\text{Im}(X)$.

On peut montrer que l'OLS a de bonnes propriétés statistiques.

Proposition 1.2. *Sous les hypothèses de la Définition 1.1, l'estimateur $\widehat{\beta}$ est sans biais, $\mathbb{E}[\widehat{\beta}] = \beta$ et de variance $\text{Var}(\widehat{\beta}) = \sigma^2(X^\top X)^{-1}$.*

Démonstration. On peut calculer son espérance en utilisant $\mathbb{E}[\varepsilon] = 0$:

$$\mathbb{E}[\widehat{\beta}] = \mathbb{E}[(X^\top X)^{-1}X^\top Y] = \mathbb{E}[(X^\top X)^{-1}X^\top X\beta + \cancel{(X^\top X)^{-1}X^\top} \varepsilon] = \beta.$$

et sa variance en utilisant $\text{Var}(Y) = \text{Var}(\varepsilon) = \sigma^2 I_n$.

$$\text{Var}(\widehat{\beta}) = \text{Var}((X^\top X)^{-1}X^\top Y) = (X^\top X)^{-1}X^\top \text{Var}(Y)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

\square

On peut même montrer que l'OLS vérifie la propriété de Gauss-Markov. Il est optimal parmi les estimateurs sans biais de β , dans le sens où il a une matrice de variance-covariance minimale (pour une certaine relation d'ordre partiel).

Cependant, la Proposition 1.2 est peu pratique pour estimer la fiabilité de $\widehat{\beta}$ en pratique car elle fait intervenir la variance du bruit, qui est inconnue. Il peut également être intéressant de l'estimer aussi. Un estimateur naturel et de regarder l'erreur moyenne

$$\widehat{\sigma}_{\text{nat}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - X_{i,\cdot} \widehat{\beta})^2 = \frac{\|Y - X \widehat{\beta}\|^2}{n}.$$

Cependant cet estimateur est biaisé. On utilise donc l'estimateur non biaisé (exercice) :

$$\widehat{\sigma}^2 = \frac{\|Y - X \widehat{\beta}\|^2}{n - p}.$$

On peut de même estimer le risque associé à l'OLS.

Proposition 1.3. Soit $(X_{n+1}, Y_{n+1}) \in \mathbb{R}^p \times \mathbb{R}$ satisfaisant le modèle linéaire $Y_{n+1} = \beta X_{n+1} + \varepsilon_{n+1}$, avec ε_{n+1} bruit centré indépendant de $\varepsilon \in \mathbb{R}^n$ et de variance $\text{Var}(\varepsilon_{n+1}) = \sigma^2$. Alors

$$\begin{aligned} \mathbb{E}[\mathcal{R}(\widehat{\beta})] &:= \mathbb{E}[(Y_{n+1} - X_{n+1}^\top \widehat{\beta})^2] \\ &= \mathbb{E}[(X_{n+1}^\top \widehat{\beta} - \mathbb{E}[X_{n+1}^\top \widehat{\beta}])^2] + \mathbb{E}[(\mathbb{E}[X_{n+1}^\top \widehat{\beta}] - \mathbb{E}[Y_{n+1}])^2] + \mathbb{E}[(Y_{n+1} - \mathbb{E}[Y_{n+1}])^2] \\ &= X_{n+1}^\top \text{Var}(\widehat{\beta}) X_{n+1} + \sigma^2 \\ &= \sigma^2 (1 + X_{n+1}^\top (X^\top X)^{-1} X_{n+1}). \end{aligned}$$

Remarque : X et X_{n+1} sont supposés déterministes et connus ici. On n'a donc pas explicité le fait que les espérances $\mathbb{E}[Y]$ et $\mathbb{E}[X^\top \beta]$ ici sont conditionnelles à X et X_{n+1} .

Le cas Gaussien Un cas particulier très considéré est celui du bruit Gaussien : $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Ce choix vient non seulement du fait qu'il permet de calculer de nombreuses propriétés statistiques supplémentaires sur $\widehat{\beta}$ et de faire des tests (intervalles de confiance, significativité des variables, ...). Il est en pratique motivé par le théorème central limite et le fait que le bruit est souvent une addition de nombreux phénomènes non expliqués par la combinaison linéaire des variables explicatives.

Proposition 1.4. Dans le cas Gaussien, les estimateurs du maximum de vraisemblance de β et σ vérifient respectivement :

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

On retrouve donc l'estimateur des moindres carrés obtenu par la minimisation du risque empirique. L'estimateur de la variance est biaisé.

Problème non-linéaire : regression polynomiale, spline, à noyau

L'hypothèse que les observations y_i s'explique comme une combinaison des variables explicatives $x_{i,j}$ peut paraître forte. Cependant, on peut appliquer la théorie précédente sur des transformations des variables $x_{i,j}$. Par exemple, en ajoutant les puissance des variables $x_{i,j}^k$ ou leurs produits $x_{i,j}x_{i,j'}$. On a alors une matrice de variable explicative X de la forme $[X_1, \dots, X_p, X_1^2, \dots, X_p^2, X_1X_2, \dots, X_1X_q, \dots]$. Cela permet de se comparer aux espaces polynomiaux.

Faire une régression linéaire sur des transformation polynomiales des variables revient à faire une regression polynomiale. Bien sûr d'autres bases/transformations existent comme des regression sur des bases de splines (polynomes par morceaux avec contraintes sur les bords). C'est le modèle qu'utilise par exemple EDF en opérationnel pour prévoir la consommation électrique en fonction de variables comme l'heure, le jour de la semaine, la température ou la couverture nuageuse. Une autre forme de régression que nous aborderons dans la suite sera la régression à noyau.

Dans la Figure 1, on a de cette façon minimisé le risque empirique sur les espaces polynomiaux de degré 1 (modèle linéaire), 3 et 30. On voit qu'il faut faire attention à ne pas considérer des espaces trop grands, au risque que le modèle soit mal posé (X non injective). Inversement, pour que la Proposition 1.3 soit vérifiée, il faut se trouver dans le vrai modèle $Y = X\beta + \varepsilon$. Il faut donc s'assurer que X contienne suffisamment de descripteurs pour que la dépendance entre Y et X soit effectivement linéaire. Autrement on paye un terme de biais supplémentaire.

Identifiabilité Si X n'est pas injective (i.e., $\text{rg}(X) = p$), la matrice $(X^\top X)$ n'est plus inversible et le problème admet plusieurs solutions qui minimisent le risque empirique. On dit que le problème est mal posé ou non identifiable.

La Proposition 1.3 nous rappelle que le risque augmente proportionnellement à la variance de $\hat{\beta}$ qui elle-même dépend du conditionnement de la matrice $(X^\top X)^{-1}$. Plus les colonnes de cette dernière risquent d'être dépendantes et moins $\hat{\beta}$ sera précis. Plusieurs solutions permettent de traiter le cas où $\text{rg}(X) < p$:

- *contrôle explicite de la complexité* en réduisant l'espace des solutions $\text{Im}(X)$. Cela peut se faire en retirant des colonnes de la matrice X jusqu'à ce qu'elle deviennent injective (par exemple, en réduisant le degré des polynômes). On peut aussi poser des contraintes d'identifiabilité de la forme $\beta \in V$ un sous-espace vectoriel de \mathbb{R}^p tel que tout élément $y \in \text{Im}(X)$ a un unique antécédent $\beta \in V$ avec $y = X\beta$. On peut par exemple choisir $V = \text{Ker}(X)^\perp$.

— *contrôle implicite de la complexité* en régularisant le problème de minimisation du risque empirique comme on le voit rapidement ci-après avec les régression Ridge et Lasso.

Regression Ridge La régularisation la plus courante est la *regression Ridge*, où on utilise la norme 2 :

$$\widehat{\beta}_{(\text{ridge})} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}.$$

Le paramètre de régularisation $\lambda > 0$ règle le compromis entre la variance de $\widehat{\beta}$ et son biais. En refaisant les calculs de la Proposition 1.1, on voit que la solution est unique (même si X n'est pas injective) donnée par :

$$\widehat{\beta}_{(\text{ridge})} = (X^\top X + n\lambda I_n)^{-1} X^\top Y.$$

On voit qu'il n'y a plus le problème d'inversion de $X^\top X$ puisque la régression Ridge revient à remplacer $(X^\top X)^{-1}$ par $(X^\top X + n\lambda I_n)^{-1}$ dans la solution de l'OLS. Le calibration du paramètre λ est cependant essentielle en pratique. Elle peut par exemple se faire de façon analytique, par *validation croisée (généralisée)* ou par *l'heuristique de pente* (cf. cours sur la sélection de modèle).

Proposition 1.5 (Biais et variance de $\widehat{\beta}_{(\text{ridge})}$). *L'estimateur Ridge vérifie :*
 $\widehat{\beta}_{(\text{ridge})} = (X^\top X + \lambda I_p)^{-1} X^\top X \beta$, $\mathbb{E}[\widehat{\beta}_{(\text{ridge})}] = \beta - \lambda (X^\top X + \lambda I_p)^{-1} \beta$ et
 $\text{Var}(\widehat{\beta}_{(\text{ridge})}) = \sigma^2 (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1}$.

L'estimateur Ridge est donc biaisé au contraire de l'OLS ce qui constitue un handicap. Cependant, sa variance ne fait pas intervenir l'inverse de $X^\top X$ mais de $X^\top X + \lambda I_p$ qui est mieux conditionnée. Il aura donc une plus faible variance. Exercice : calculer le risque de l'estimateur Ridge.

Le Lasso Une deuxième régularisation courante est la régularisation par la norme 1. On parle alors de l'estimateur Lasso.

$$\widehat{\beta}_{(\text{lasso})} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

L'estimateur Lasso n'a pas de formule analytique dans le cadre général. Cependant, dans le cas où $X^\top X = I_p$, le Lasso a une solution explicite. L'estimateur Lasso correspond alors à un seuillage doux de $\widehat{\beta}$ la solution des moindres carrés (OLS) : pour tout $j = 1, \dots, p$

$$\widehat{\beta}_{\text{lasso}} = \text{sgn}(\widehat{\beta}) \max(0, |\widehat{\beta}| - \lambda).$$

L'estimateur Lasso est en particulier très utilisé quand le nombre de variables explicatives est très grand voir supérieur au nombre d'observations : $p \gg n$. Le caractère "anguleux" de la norme 1 va avoir tendance à tronquer (annuler) de nombreux coefficients de $\widehat{\beta}_{\text{lasso}}$ pour proposer des solutions parcimonieuses. Ceci se voit mieux avec la formulation équivalente du Lasso sous la forme d'un problème de régularisation sous contrainte : pour chaque $\lambda > 0$, il existe $U = U(X) > 0$ tel que :

$$\widehat{\beta}_{(\text{lasso})} \in \arg \min_{\beta \in \mathbb{R}^p: \|\beta\|_1 \leq U} \|Y - X\beta\|^2.$$

Ridge se reformule de la même façon en remplaçant la norme 1 par la norme 2 au carré.

Calcul de $\widehat{\beta}$: décomposition QR et descente de gradient. La formule close $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$ de l'OLS est utile pour l'analyser. Cependant, la calculer de façon naïve peut s'avérer prohibitif. En particulier quand p est grand, on préfère éviter l'inversion de la matrice $X^\top X$ qui coûte $\mathcal{O}(p^3)$ par la méthode de Gauss-Jordan et qui peut être très instable quand la matrice est mal conditionnée.

Cependant, pour améliorer la stabilité, on peut utiliser la décomposition QR . Rappelons que $\widehat{\beta}$ est solution de l'équation :

$$X^\top X \widehat{\beta} = X^\top Y.$$

On écrit $X \in \mathbb{R}^{n \times p}$ de la forme $X = QR$, où $Q \in \mathbb{R}^{n \times p}$ est une matrice orthogonale (ie, $QQ^\top = I_n$) et $R \in \mathbb{R}^{p \times p}$ triangulaire supérieure. Les matrices triangulaires supérieures sont très utiles pour résoudre les systèmes linéaires. En substituant dans l'équation précédente, on obtient :

$$\begin{aligned} R^\top (Q^\top Q) R \widehat{\beta} &= R^\top Q^\top Y \Leftrightarrow R^\top R \widehat{\beta} = R^\top Q^\top Y \\ &\Leftrightarrow R \widehat{\beta} = Q^\top Y. \end{aligned}$$

Il ne reste alors plus qu'à résoudre un système linéaire avec une matrice triangulaire supérieure, ce qui est facile.

Pour éviter totalement le coût de l'inversion de matrice, une autre solution est la descente de gradient. Qui consiste à résoudre le problème de minimisation pas à pas en s'approchant du minimum grâce à des pas de gradient. On initialise par exemple $\widehat{\beta}_0 = 0$, puis on met à jour :

$$\begin{aligned} \widehat{\beta}_{i+1} &= \widehat{\beta}_i - \eta \nabla \widehat{\mathcal{R}}_n(\widehat{\beta}_i) \\ &= \widehat{\beta}_i - \frac{2\eta}{n} ((X^\top X) \widehat{\beta}_i - Y^\top X), \end{aligned}$$

où $\eta > 0$ est un paramètre d'apprentissage. On voit que si l'algorithme converge, alors il converge vers un point annulant le gradient, donc vers la solution de l'OLS. Pour avoir convergence, il faut que le paramètre η soit bien calibré, mais on verra cela plus en détails dans le cours sur la descente de gradient.

Si le jeu de données est beaucoup trop grand $n \gg 1$. Il peut également être prohibitif de charger toutes les données pour faire le calcul de $\nabla \widehat{\mathcal{R}}_n(\widehat{\theta}_i)$. La solution courante est alors de faire une descente de gradient stochastique, où on ne fait des pas de gradients que sur des estimés de $\nabla \widehat{\mathcal{R}}_n(\widehat{\theta}_i)$, calculés sur une sous-partie aléatoire des données.

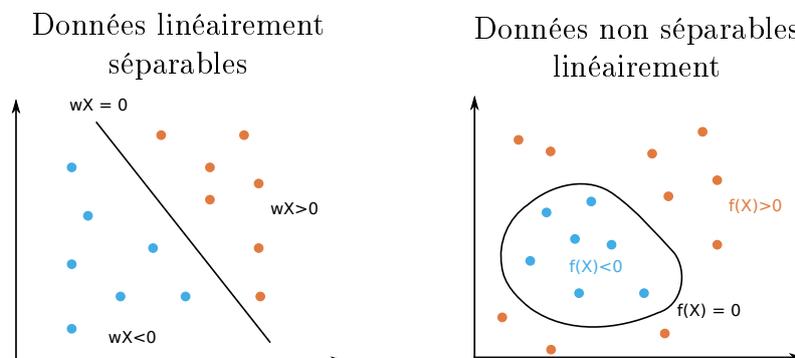
2 Regression logistique

On va maintenant s'intéresser au problème de classification. Dans celui-ci, on veut associer à toute observation $x \in \mathbb{R}^d$ un étiquette $y \in \{0, 1, \dots, M\}$. Dans ce qui suit, on se restreint au cas $M = 1$ correspondant à la classification binaire. Le but de la régression logistique est d'étendre l'OLS à la classification.

On reprend les notations du modèle linéaire (partie 1.1) de $Y \in \mathbb{R}^n$, la matrice de variables explicatives $X \in \mathbb{R}^{n \times p}$. La différence est qu'on suppose que les observations Y sont binaires et appartiennent à $\{0, 1\}^n$. On ne peut plus expliquer les y_i par des combinaisons linéaires des $x_{i,j}$ puisque ces dernières appartiennent à \mathbb{R} et non pas $\{0, 1\}$. On va donc chercher ici à classifier les données en proposant $f : \mathbb{R}^n \mapsto \mathbb{R}$ tel que :

$$f(x_i) \begin{cases} \geq 0 & \Rightarrow y_i = +1 \\ < 0 & \Rightarrow y_i = 0 \end{cases},$$

afin de séparer les données en deux groupes. En particulier, on va s'intéresser au cas où f est linéaire de la forme $\beta \mapsto X\beta$ ce qui suppose que les données sont bien expliquées par une séparation linéaire.



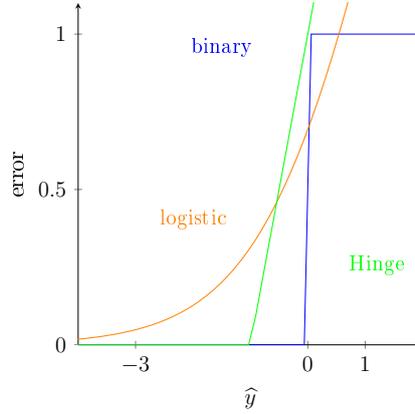


FIGURE 2 – Binary, logistic and Hinge loss incurred for a prediction $\hat{y} := x^\top \beta$ when the true observation is $y = 0$.

Bien sûr, si les données ne semblent pas linéairement séparables on peut utiliser la même astuce que pour la régression polynomiale. On essaie de se placer dans un espace de plus grande dimension, en ajoutant des transformations des variables, dans lequel les données sont linéairement séparables.

Pour pouvoir minimiser le risque empirique, il reste à déterminer la fonction de perte. Une perte naturelle est la perte binaire : on perd 1 si on s'est trompé et 0 sinon. Le risque empirique s'exprime alors :

$$\widehat{\mathcal{R}}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i \neq \mathbb{1}_{x_i^\top \beta \geq 0}}.$$

Cette perte n'est cependant pas convexe en une prévision $x_i^\top \beta$, ce qui rend la minimisation de $\widehat{\mathcal{R}}_n$ très compliquée. Une idée consiste à remplacer la perte par une similaire et convexe. C'est le cas de la perte Hinge, mais aussi de la perte logistique $\ell : \{0, 1\} \times \mathbb{R} \rightarrow \mathbb{R}_+$, qui à une prévision $\hat{y} = x^\top \beta$ et une observation $y \in \{0, 1\}$ associe

$$\ell(\hat{y}, y) = y \log(1 + e^{-\hat{y}}) + (1 - y) \log(1 + e^{\hat{y}}). \quad (4)$$

Les pertes binaires, Hinge et logistiques sont représentées en Figure 2.

Definition 2.1 (Estimateur de regression logistique). *L'estimateur de la regression logistique est la solution du problème de minimisation :*

$$\widehat{\beta}_{(\text{logit})} = \frac{1}{n} \sum_{i=1}^n \ell(x_i^\top \beta, y_i),$$

où ℓ est la perte logistique définie en équation (4).

Interprétation probabiliste L'avantage de la perte logistique par rapport à la perte Hinge est qu'elle a une interprétation probabiliste en modélisant $P(Y = 1|X)$, où ici (X, Y) est un couple de variables aléatoires suivant la loi des observation (x_i, y_i) . Par la loi de Bayes, nous avons :

$$\begin{aligned}\mathbb{P}(Y = 1|X) &= \frac{\mathbb{P}(X|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X|Y = 1)\mathbb{P}(Y = 1) + \mathbb{P}(X|Y = 0)\mathbb{P}(Y = 0)} \\ &= \frac{1}{1 + \frac{\mathbb{P}(Y=0)\mathbb{P}(X|Y=0)}{\mathbb{P}(Y=1)\mathbb{P}(X|Y=1)}}.\end{aligned}$$

On note

$$f(X) := \log \left(\frac{\mathbb{P}(Y = 1)\mathbb{P}(X|Y = 1)}{\mathbb{P}(Y = 0)\mathbb{P}(X|Y = 0)} \right) = \log \left(\frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)} \right) + \log \left(\frac{\mathbb{P}(X|Y = 1)}{\mathbb{P}(X|Y = 0)} \right)$$

le log de rapport de chance entre la classe 1 et la classe 0. Cette fonction caractérise complètement la probabilité postérieure $\mathbb{P}(Y = 1|X)$. On a donc :

$$\mathbb{P}(Y = 1|X) = \frac{1}{1 + e^{-f(X)}} =: \sigma(f(X)) \quad \text{avec} \quad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

La fonction σ est appelée fonction logistique et satisfait $\sigma(-z) = 1 - \sigma(z)$ et $\frac{d\sigma(z)}{dz} = \sigma(z)\sigma(-z)$. Son intérêt est qu'elle permet de transformer une fonction f à valeur dans \mathbb{R} en une probabilité entre 0 et 1.

La régression logistique revient en fait à supposer que f est linéaire de la forme $f : x \mapsto x^\top \beta$. On rappelle que comme pour la régression linéaire, pour simplifier les notation, nous obsorbons une ordonnée à l'origine en augmentant les variables explicatives de la constante 1. L

Proposition 2.1. *En supposant, que $(x_i, y_i)_{1 \leq i \leq n}$ est un n -échantillon tel que $\mathbb{P}(y_i = 1|x_i) = \sigma(x_i^\top \beta)$, alors l'estimateur du maximum de β est $\hat{\beta}_{(\text{logit})}$.*

Démonstration. La log-vraisemblance s'écrit :

$$\begin{aligned}\log L &= \sum_{i=1}^n \log (\mathbb{P}_\beta(y_i = 1|x_i)^{y_i} (1 - \mathbb{P}_\beta(y_i = 0|x_i))^{1-y_i}) \\ &= \sum_{i=1}^n \log (\sigma(\beta^\top x_i)^{y_i} \sigma(-\beta^\top x_i)^{1-y_i}) \\ &= \sum_{i=1}^n \ell(\beta^\top x_i, y_i)\end{aligned}$$

où ℓ est la perte logistique. L'estimateur de la regression logistique est donc le minimiseur de la vraisemblance. \square

Calcul de $\widehat{\beta}_{(\text{logit})}$ Comme pour l'OLS, on peut tenter de résoudre le problème analytiquement en annulant le gradient du risque empirique associé. Comme $\frac{\partial \ell(\widehat{y}, y)}{\partial \widehat{y}} = \sigma(\widehat{y}) - y$, on a :

$$\nabla \widehat{\mathcal{R}}_n(\beta) = \frac{1}{n} \sum_{i=1}^n x_i (\sigma(x_i^\top \beta) - y_i) = \frac{1}{n} X (Y - \mu(\beta))$$

où $\mu(\beta)_i := \sigma(x_i^\top \beta)$ pour $1 \leq i \leq n$. Mauvaise nouvelle, l'équation $\nabla \widehat{\mathcal{R}}_n(\beta)$ ne se résoud pas sous forme analytique. Il faut donc avoir recours à des algorithmes itératifs (descente de gradient, méthode de Newton, ...). C'est heureusement possible car la perte logistique est convexe en \widehat{y} , comme on le voit en calculant sa dérivée seconde :

$$\frac{\partial^2 \ell(\widehat{y}, y)}{\partial \widehat{y}^2} = \sigma(\widehat{y})\sigma(-\widehat{y}) > 0.$$

La dérivée seconde étant même fortement convexe, le problème aura une unique solution.

Newton pour la régression logistique : moindres carrés pondérés itérés Si il y a le temps...

Régularisation La régression logistique a les mêmes risques de sur-apprentissage que la régression linéaire (en particulier quand $p > n$). Il faut donc régulariser en ajoutant un terme par exemple $\lambda \|\beta\|_2^2$ à la perte logistique.

Références

- CORNILLON, P.-A. et E. MATZNER-LØBER (2011). “La régression linéaire simple”. In : *Régression avec R*, p. 1–28.
- RIVOIRARD, V. et G. STOLTZ (2012). *Statistique mathématique en action*.