



Modélisation Linéaire Licence 3 MIASHS et Informatique


Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

Ce cours est à destination des étudiants de la Licence 3 MIASHS et Informatique de l'Université Lumière, Lyon 2. Il se présente en deux parties :

Les différentes méthodes d'analyses seront illustrées avec des exemple sous .

Il faut rédiger une partie sur les vecteurs gaussiens et Cochran

Faire l'ANOVA et terminer la régression logistique (développer un peu plus la partie mesures de performances qui est trop succincte pour le moment) + écrire les critères permettant de faire de la sélection de modèle -> Ecrire la fin de ce cours!!! Rédiger la partie estimation non paramétrique -> Nadaraya Watson et préparer les code R associés. Penser aussi à traiter les modèles mixtes et les modèle à effets mixtes. -> Pas cette année je pense ... à voir.


Remerciements Un grand merci à Stéphane Chrétien dont les ressources ont été d'une aide précieuse pour la préparation de ce cours tant sur le contenu que sur la création des fiches d'exercices et de séances pratiques. Merci également à Francesco Amato pour m'avoir grandement aidé dans la rédaction des corrections des fiches de TD.

Un grand merci également aux différents chargés de TD qui ont permis d'améliorer les exercices et les corrections de ces derniers à travers les différentes remarques constructives :

- 2023-2024 : Francesco Amato et Alejandro Rivera
- 2023-2024 : Francesco Amato et Zied Gharbi

Table des matières

I	Introduction	5
1	Dérivation des fonctions à plusieurs variables	9
1.1	Fonctions différentiables d'ordre 1	9
1.2	Fonctions différentiables d'ordre 2	18
2	Recherche d'extrema et applications aux fonctions convexes	23
2.1	Retour sur la convexité	23
2.2	Conditions d'optimalité	24
II	Modèles Linéaires Gaussiens	32
3	Modèle Linéaire Gaussien simple	34
3.1	Hypothèse du modèle Gaussien	34
3.2	Optimisation	35
3.3	Expression des solutions	36
3.4	Propriétés des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$	39
3.5	Estimation de la variance σ^2	42
3.6	Mesure du lien entre la variable explicative et la variable à expliquer	43
3.7	Lien entre la pente du modèle $\hat{\beta}_1$ et le coefficient de corrélation $\hat{\rho}$	43
3.8	Significativité du modèle	44
3.9	Ecriture du modèle sous forme matricielle	46
4	Vecteur gaussien et géométrie des modèles linéaires	47
4.1	Vecteurs gaussiens	47
4.2	Application à la variance	49
4.3	Définition des lois de probabilités	50

5	Modèle Linéaire Gaussien Multiple	52
5.1	Estimation par la méthode des moindres carrés	52
5.2	Estimation par maximum de vraisemblance	57
5.3	Estimation de la variance σ^2	60
5.4	Test de nullité d'un coefficient de la régression	61
5.5	Prédictions et intervalles de prédiction	64
5.5.1	Intervalle de confiance sur la prédiction	64
5.5.2	Intervalle de confiance sur la valeur moyenne prédite	65
5.6	Qualité du modèle	66
5.7	Construction et sélection de modèles	70
5.8	Analyse des résidus et détection d'outliers	73
6	Régression linéaire avec 	77
III	Modèles Linéaires Généralisés	78
6.1	Vers l'intérêt de la régression logistique	78
7	Vers la Régression Logistique	79
8	Régression à noyaux	83
9	Modèles (à effets) Mixtes	84
	Bibliographie	85

Première partie

Introduction

La modélisation linéaire (ou non linéaire) a pour but de chercher à décrire des phénomènes à l'aide d'une équation liant des variables aléatoires. Plus précisément, on va chercher à prédire ou expliquer les valeurs d'une variable aléatoire Y à l'aide de plusieurs variables explicatives X_1, X_2, \dots, X_p .

Pour établir ce lien, nous nous basons sur des observations qui vont nous permettre d'estimer les paramètres du modèle décrivant le phénomène. Cependant, le processus de récolte ou de traitement des données est sujet aux erreurs ce qui peut induire un biais dans le modèle appris. Cette erreur est souvent modélisée par une variable aléatoire ε dont la nature dépendra du type de modèle considéré.

Finalement, la modélisation consistera à déterminer la fonction inconnue f qui va permettre de lier la variable à expliquer Y aux variables explicatives X_1, \dots, X_p en tenant compte d'un éventuel bruit (notre erreur) dans les données, *i.e.*,

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

où $\mathbf{X} = (X_1, X_2, \dots, X_p)$ et f est la fonction que l'on cherche à déterminer et qui va dépendre de paramètres.

Ce travail de modélisation est souvent accompagné d'une phase exploratoire des données.

Statistiques

Exploration + Modélisation \longrightarrow *Data Mining*

Quelques problématiques

La nature de la modélisation change en fonction de la nature de Y :

- si Y est qualitative, on parle alors d'un problème de **classification**

- si Y est quantitative, on parlera de problème de **régression**.

Ce sont des contextes très classiques en modélisation. Il existe un dernier cas, non traité ici qui correspond au cas où l'on ne dispose pas de variable Y , mais uniquement des variables explicatives et on cherche à construire des groupes. On parle alors de **clustering**, une méthode souvent appliquée dans la *statistique en grande dimension*.

Choix du modèle

Il existe plusieurs façons d'estimer la fonction f qui peuvent aboutir à différentes estimations. Ces dernières peuvent également dépendre de la quantité d'informations utilisée, *i.e.* du nombre de variables explicatives que l'on va employer. L'objectif des problèmes de régression que nous étudierons sera de trouver un équilibre entre :

- **un nombre important de variables explicatives** : ce qui permettra au modèle de mieux expliquer les données mais avec un pouvoir prédictif plus faible, *i.e.*, un risque de mauvaises prédictions plus élevé.
- **peu de variables explicatives** : le modèle présentera une faible variance, donc des prédictions potentiellement plus faible. En revanche ce dernier aura plus de difficultés à expliquer les données.

Cette notion fera ultérieurement référence au **compromis biais - variance du modèle**, une notion très importante que vous retrouverez dans un contexte d'apprentissage statistique et qui sera liée à la notion de **complexité du modèle**. Cette dernière étant très liée à la quantité d'information, donc de variables employées.

Technique de sélections de modèles

Ces dernières sont souvent classées en deux catégories

Sélection de variables

Elle repose sur des **critères statistiques** de mesure qualité d'un modèle qui tiennent compte du nombre de paramètres employés.

Des tests statistiques sont ensuite employés pour déterminer si la différence de résultats est significative ou non.

Régularisation ou pénalisation

Procédé souvent utilisé dans la statistiques en grande dimension pour sélectionner de façon automatique les variables les plus pertinentes.

Cela se fait en ajoutant des termes dits de **pénalités** au sein du problème que l'on cherche à résoudre.

Modèles étudiés.

Nous nous intéresserons à plusieurs modèles dans le cadre de ce cours. Le plus simple étant le modèle linéaire gaussien se présentant sous la forme

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

où nous ferons l'hypothèse que \mathbf{X} sont déterministes et que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, où σ^2 désigne la variance ou le bruit présent dans nos données.

La fonction f aura alors une forme très spécifique dans ce contexte. Nous prendrons une fonction affine de la forme

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p = \sum_{j=0}^p \beta_j x_j.$$

Nous verrons également, à travers la régression logistique, comment traiter un problème de **classification** avec un modèle dit de **régression**. Dans ce contexte là, la variable que l'on cherchera à expliquer

Nous verrons également comment l'ANOVA, présentée en statistiques inférentielles peut être abordée selon un point de vue différents à l'aide de cette théorie.

Enfin, il sera aussi intéressant de voir que les problèmes de régressions peuvent aussi être étudiés et développées avec des méthodes non paramétriques, comme la **régression à noyaux** ou encore par des méthodes plus informatiques à l'aide de **réseaux de neurones**.

Pour l'ensemble de ces modèles, nous en étudierons les propriétés statistiques et nous verrons les tests qui permettent de montrer la pertinence ou non des informations utilisées.

Les notions abordées lors du cours de statistiques inférentielles pourront également nous servir à construire des intervalles de confiance sur les estimateurs mais aussi sur les prédictions effectuées par le modèle.

Nous attacherons également une grande importance à l'interprétation géométrique des modèles à l'aide des outils d'algèbre linéaire.

Pré-requis.

Pour ce cours, il sera essentiel d'être à l'aise avec le calcul matriciel, la manipulation des vecteurs ou encore la notion de projection sur des sous-espaces.

Il est donc vivement conseillé de se reporter au cours disponible à l'[adresse suivante](#).

Il sera également intéressant de connaître des éléments de base en calcul différentiel, *i.e.*, sur la dérivation de fonctions à plusieurs variables ainsi que la caractérisation de la convexité. Une source présentant ces différentes notions est disponible à l'[adresse suivante](#). On fera cependant un rappel de ces différentes notions d'analyse dans les sections qui suivent.

1 Dérivation des fonctions à plusieurs variables

On revient sur les notions fondamentales de dérivabilité des fonctions à plusieurs et plus précisément sur les notions de fonctions différentiables à l'ordre 1 et à l'ordre 2.

1.1 Fonctions différentiables d'ordre 1

Le nombre dérivée d'une fonction réelle à valeurs réelles f en un point x_0 a été définie comme décrivant l'évolution de la fonction f au voisinage de ce point. Plus précisément, nous l'avons défini comme la variation de f au voisinage de x_0 par la quantité

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Dans le cas unidimensionnel, nous n'avons que deux façons de faire tendre h vers 0, par valeurs inférieures ou valeurs supérieures. Dans le cas multi-dimensionnel il ne faudra plus considérer des h réels mais vectoriels car il existe une infinité de directions qui pointent vers un point \mathbf{x}_0 , ce qui nous ramène à notre étude des espaces vectoriels et la décomposition des vecteurs dans une base.

La direction \mathbf{h} considérée étant à présent un vecteur de \mathbb{R}^n , on peut donc le décomposer dans une base, la base canonique $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ par exemple et écrire

$$\mathbf{h} = h_1 \mathbf{e}_1 + h_2 \mathbf{e}_2 + \dots + h_n \mathbf{e}_n.$$

Ainsi, en étudiant les variations rapport aux différents vecteurs de la base, on pourra étudier les variations de la fonctions par rapport à n'importe quel vecteur \mathbf{h} et donc dans n'importe quelle direction.

Ce dernier point laisse donc à penser que la dérivée, que l'on l'appellera **gradient**, d'une fonction à plusieurs variables sera donc un **vecteur**.

De cette remarque, on va donc s'intéresser aux **dérivées partielles** de la fonction par à ces différentes composantes.

Définition 1.1: Fonctions partielles

Soit f une fonction définie sur un domaine de D de \mathbb{R}^n et $\mathbf{x} \in D$.

Pour tout $i \in \llbracket 1, n \rrbracket$, on appelle $f_{\mathbf{x}}^i$ les fonctions partielles définies par

$$f_{\mathbf{x}}^i(u) = f(x_1, x_2, \dots, x_{i-1}, u, x_{i+1}, \dots, x_n),$$

où u est tel que $\{(x_1, x_2, \dots, x_{i-1}, u, x_{i+1}, \dots, x_n)\} \subset D$.

Une fonction partielle peut finalement être vue comme une fonction d'une seule variable, *i.e.*, une fonction pour laquelle *toutes les variables sont figées sauf une*.

Dans le cas où ces différentes fonctions partielles sont dérivables, on parle alors de dérivées partielles.

Définition 1.2: Dérivées partielles

Soit f une fonction définie d'un ensemble D de \mathbb{R}^n dans \mathbb{R} . Si la fonction f admet des dérivées partielles en tout point $\mathbf{x} \in D$, alors on note $\frac{\partial f}{\partial x_i}$ l'application

$$\mathbf{x} \mapsto \frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{\partial f(x_1, x_2, \dots, x_i, \dots, x_n)}{\partial x_i}(\mathbf{x}).$$

On peut facilement faire une analogie avec la définition de nombre dérivée pour les fonctions d'une variable réelle

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}.$$

On regardera seulement la variation dans une direction, ou selon une composante, du vecteur \mathbf{x} .

Exemple 1.1. Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction affine définie par $f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + c$, où $\mathbf{a} \in \mathbb{R}^n$ et $c \in \mathbb{R}$.

Alors les dérivées partielles de f existent et sont définies, pour tout $i \in \llbracket 1, n \rrbracket$ par

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \mathbf{a}$$

Exemple 1.2. Considérons la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x, y) = e^{-x^2+y^2} - 3xy + \frac{1}{x^2 + y^2 + 1}$.

La fonction f admet des dérivées partielles par rapport à x et à y qui sont données par

$$\frac{\partial f}{\partial x}(x, y) = -2xe^{-x^2+y^2} - 3y - \frac{2x}{(x^2 + y^2 + 1)^2} \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = -2ye^{-x^2+y^2} - 3x - \frac{2y}{x^2 + y^2 + 1}.$$

On peut également définir le concept de fonctions de classe C^1 pour les fonctions à plusieurs variables, il faut et il suffit que les différentes dérivées partielles soient continues.

Cependant, la généralisation à des fonctions de classe C^n n'est pas aussi évidente. nous limiterons au cas des fonctions de classe C^2 que nous étudierons ultérieurement.

Maintenant que nous avons introduit les dérivées partielles, on peut introduire la notion de **gradient** qui représente la valeur de la dérivée d'une fonction f selon toutes ses composantes.

Définition 1.3: Gradient

Soit f une fonction définie sur un domaine D de \mathbb{R}^n à valeurs dans \mathbb{R} et soit \mathbf{a} un point de D . Si f admet des dérivées partielles d'ordre 1 par rapport aux différentes variables, alors, on appelle **gradient de f en \mathbf{a}** et on note $\nabla f(\mathbf{a})$, le vecteur de \mathbb{R}^n définie par

$$\nabla f(\mathbf{a}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{a}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{a}) \end{pmatrix}$$

De plus, si f est une fonction de classe C^1 , alors le gradient est une application continue de D dans \mathbb{R}^n .

Exemple 1.3. Considérons la fonction $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ définie par

$$f(x, y, z) = 3x^2yz - 2x^3y^2z^4 + e^x.$$

La fonction f admet des dérivées partielles continues en tout point de D et son gradient est donnée par

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 6xyz - 6(xy z^2)^2 + e^x \\ 3x^2z - 4x^3y z^4 \\ 3x^2y - 8x^3y^2z^3 \end{pmatrix}.$$

Lorsque la fonction est à valeurs vectorielles la notion de gradient change car l'image n'est plus un simple nombre réel mais un vecteur. Il sera donc important de regarder comment se comporte les différentes composantes de l'image par la fonction selon les différentes variables.

Exemple 1.4. Considérons la fonction $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ définie par

$$f(x, y, z) = (3xy, 6xy, -2yz).$$

On voit que chaque composante évolue différemment selon les variables x, y et z . On va donc regarder chaque composante de l'image une par une et leurs différentes dérivées partielles.

On sera donc amenés à calculer un nombre de dérivées partielles qui dépend à la fois du nombre de variables, mais aussi de la dimension de l'espace d'arrivée. On pourra alors obtenir, dans le cas d'une fonction à valeurs vectorielles, une **matrice**, que l'on appelle la **matrice Jacobienne**.

Définition 1.4: Matrice Jacobienne

Soit $\mathbf{a} \in \mathbb{R}^n$ et soit f une fonction définie sur un ensemble D de \mathbb{R}^n dans \mathbb{R}^p telle que

$$f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_p(\mathbf{x})),$$

où les fonctions f_j sont des fonctions à valeurs réelles qui admettent des dérivées partielles d'ordre 1.

On appelle **matrice jacobienne** de f en \mathbf{a} , notée $Jac_f(\mathbf{a}) \in \mathbb{R}^{p \times n}$ définie par

$$Jac_f(\mathbf{a}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \vdots & & \vdots \\ \frac{\partial f_p}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_p}{\partial x_n}(\mathbf{a}) \end{pmatrix}$$

La matrice Jacobienne est alors une généralisation gradient pour les fonctions à valeurs réelles.

Exemple 1.5. Soit $\mathbf{A} \in \mathbb{R}^{p \times n}$, soit $f : \mathbb{R}^n \rightarrow \mathbb{R}^p$ définie par $f(\mathbf{x}) = \mathbf{Ax}$.

Alors le Jacobien de f est donné par la matrice \mathbf{A} elle même.

De façon générale, la dérivée d'une application linéaire est elle même.

Exemple 1.6. Considérons la fonction $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ définie par

$$f(x, y, z) = (3xy, 6xy, -2yz).$$

alors la matrice Jacobienne de f en un point (x, y, z) de \mathbb{R}^3 est donnée par

$$Jac_f(\mathbf{x}) = \begin{pmatrix} 3y & 3x & 0 \\ 6y & 6x & 0 \\ 0 & -2z & -2y \end{pmatrix}$$

Jusqu'à présent, nous avons étudié les dérivées de d'une fonction différentiable f par rapport au vecteur de la base de \mathbb{R}^n . C'est-à-dire que nous avons étudié l'évolution

de la fonction f par rapport aux vecteurs de bases, *i.e.* dans n directions différentes.

A partir de cela, nous sommes maintenant capables d'étudier la dérivée de cette même fonction f dans n'importe quelle direction $\mathbf{d} \in \mathbb{R}^n$ qui va s'écrire

$$\mathbf{d} = \sum_{j=1}^n d_j \mathbf{e}_j.$$

On parle alors de **dérivée directionnelle**.

Définition 1.5: Dérivée directionnelle

Soit f une fonction définie sur un domaine $D \subset \mathbb{R}^n$, \mathbf{x} un point de D et \mathbf{d} un vecteur de norme unitaire^a. Considérons également la fonction $\phi : \mathbb{R} \rightarrow \mathbb{R}$ par $\phi(t) = f(\mathbf{x} + t\mathbf{d})$.

Si ϕ est dérivable en 0, alors f est dérivable en \mathbf{x} dans la direction \mathbf{d} et sa dérivée est noté $D_{\mathbf{d}}f(\mathbf{x})$

$$\begin{aligned} D_{\mathbf{d}}f(\mathbf{x}) &= \phi'(0), \\ &= \lim_{t \rightarrow 0} \frac{\phi(t) - \phi(0)}{t}, \\ &= \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}. \end{aligned}$$

^a. Pour rappel, cela signifie que $\|\mathbf{d}\| = \sum_{j=1}^n d_j^2 = 1$.

Dans le cas où \mathbf{d} n'est rien d'autre qu'un vecteur de base, on retrouve la définition de dérivée partielle.

Proposition 1.1: Dérivée directionnelle

Soit D un sous ensemble de \mathbb{R}^n et soit $f : D \rightarrow \mathbb{R}$ une fonction de classe C^1 sur D , \mathbf{x} un point de D et \mathbf{d} un vecteur unitaire de \mathbb{R}^n .

Alors la fonction f possède une dérivée dans la direction \mathbf{d} égale à

$$D_{\mathbf{d}}f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.$$

Démonstration. Repartons de la définition et considérons $\phi(t) = f(\mathbf{x} + t\mathbf{d}) = f(g(t))$ où $g : \mathbb{R} \rightarrow D$ définie par $g(t) = \mathbf{x} + t\mathbf{d}$. On réécrit ainsi la fonction ϕ comme une composée de deux fonctions. En appliquant la Proposition ??, nous avons

$$\phi'(t) = (f \circ g)'(t) = \langle \nabla f(g(t)), J_g(t) \rangle.$$

En $t = 0$, nous avons :

$$\phi'(0) = \langle \nabla f(g(0)), J_g(0) \rangle = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.$$

□

Regardons un petit exemple avec une fonction affine.

Exemple 1.7. Soit $\mathbf{a} \in \mathbb{R}^n$ et c un nombre réel. Considérons f la fonction affine définie de \mathbb{R}^n dans \mathbb{R} par

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + c = a_1x_1 + \dots + a_nx_n + c.$$

Considérons maintenant une direction $\mathbf{d} = (d_1, d_2, \dots, d_n)$ de \mathbb{R}^n . Alors, pour tout réel t

$$\begin{aligned} \phi(h) &= f(\mathbf{x} + t\mathbf{d}), \\ &= \langle \mathbf{a}, \mathbf{x} + t\mathbf{d} \rangle + c, \\ &= a_1(x_1 + td_1) + a_2(x_2 + td_2) + \dots + a_n(x_n + td_n). \end{aligned}$$

Donc la dérivée directionnelle est donnée par

$$a_1d_1 + \dots + a_nd_n = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.$$

Ces notions de dérivées, de gradient, nous servent principalement, comme en dimension 1 à déterminer une approximation de notre fonction au voisinage d'un point. Plus précisément, pour le moment, une approximation d'ordre 1 soit une approximation affine.

Cela suppose de deux choses. Si on note T l'approximation affine de la fonction f en un point \mathbf{a} , alors T doit vérifier

- Il faut que la fonction et son approximation coïncide en le point \mathbf{a} où la fonction est approchée

$$f(\mathbf{a}) = T(\mathbf{a}).$$

- Il faut aussi T devienne aussi proche de f lorsque \mathbf{x} tend \mathbf{a} et cela plus rapidement que la vitesse de convergence de \mathbf{x} vers \mathbf{a} , ce que l'on peut traduire par

$$\lim_{\mathbf{x} \rightarrow \mathbf{a}} \frac{f(\mathbf{x}) - T(\mathbf{x})}{\|\mathbf{x} - \mathbf{a}\|} = 0.$$

Pour obtenir une expression de cette approximation, nous avons besoin d'introduire une dernière définition, qui est celle de **différentiabilité**, même si le terme est déjà apparu auparavant. Cette définition va faire apparaître une fonction linéaire qui sera importante dans la définition de notre approximation T .

Définition 1.6: Différentielle des fonctions à plusieurs variables

Soit D un sous-ensemble de \mathbb{R}^n , f une application de D à valeurs réelles et \mathbf{a} un point de D . On dit que f est différentiable en \mathbf{a} s'il existe une fonction linéaire L telle que

$$\lim_{\mathbf{x} \rightarrow \mathbf{a}} \frac{f(\mathbf{x}) - (f(\mathbf{a}) + L(\mathbf{x} - \mathbf{a}))}{\|\mathbf{x} - \mathbf{a}\|} = 0.$$

Comme la fonction L est linéaire, elle peut s'écrire

$$L(\mathbf{x}) = \langle \mathbf{u}, \mathbf{x} \rangle = u_1 x_1 + u_2 x_2 + \dots + u_n x_n.$$

On doit donc avoir

$$\lim_{\mathbf{x} \rightarrow \mathbf{a}} \frac{f(\mathbf{x}) - (f(\mathbf{a}) + \langle \mathbf{u}, \mathbf{x} - \mathbf{a} \rangle)}{\|\mathbf{x} - \mathbf{a}\|} = 0.$$

On peut même montrer que le vecteur \mathbf{u} dont il fait mention ici n'est rien d'autre que le vecteur des dérivées partielles de la fonction f évaluées en \mathbf{a} , *i.e.*,

$$\forall i \in \llbracket 1, n \rrbracket, \quad u_i = \frac{\partial f}{\partial x_i}(\mathbf{a}).$$

Ainsi, notre fonction T est définie par

$$T(\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{a})(x_i - a_i).$$

C'est cette dernière quantité

$$\sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{a})(h) = \frac{\partial f}{\partial x_1}(\mathbf{a})h_1 + \dots + \frac{\partial f}{\partial x_n}(\mathbf{a})h_n,$$

que l'on appelle différentielle de f en \mathbf{a} évaluée en \mathbf{h} .

On remarque de suite que cette différentielle est aussi égale à

$$\langle \nabla f(a), \mathbf{h} \rangle = D_{\mathbf{h}} f(\mathbf{a}).$$

Prenons quelques exemples pour mieux comprendre les objets que l'on manipule.

Exemple 1.8. Considérons la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x_1, x_2) = x_1^2 - x_2^2$.

Et considérons le vecteur \mathbf{d} de \mathbb{R}^2 , $\mathbf{d} = (1/2, \sqrt{3}/2)$, qui est bien unitaire.

Alors le gradient de la fonction f est donnée par

$$\nabla f(\mathbf{x})^\top = (2x_1 \quad -2x_2).$$

Ainsi, la dérivée directionnelle de f dans la direction \mathbf{d} , qui est aussi la différentielle de f en \mathbf{x} évaluée en \mathbf{d} , est donnée par

$$D_{\mathbf{d}} f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle = (2x_1 \quad -2x_2) \begin{pmatrix} \frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix} = x_1 - \sqrt{3}x_2.$$

Ainsi, on peut évaluer l'accroissement de la fonction f en n'importe quel point \mathbf{x} dans la direction \mathbf{d} donnée.

Exemple 1.9. Considérons la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x, y) = \cos\left(\frac{x_1^2}{2}\right) + x_2^2$.

Et considérons le vecteur de \mathbb{R}^2 $\mathbf{d} = (4/5, 3/5)$, qui est bien unitaire.

Alors le gradient de la fonction f est donnée par

$$\nabla f(\mathbf{x})^\top = \left(-x_1 \sin\left(\frac{x_1^2}{2}\right) \quad 2x_2 \right).$$

Ainsi, la dérivée directionnelle de f dans la direction \mathbf{d} , qui est aussi la différentielle de f en \mathbf{x} évaluée en \mathbf{d} , est donnée par

$$D_{\mathbf{d}} f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle = \left(-x_1 \sin\left(\frac{x_1^2}{2}\right) \quad 2x_2 \right) \begin{pmatrix} \frac{4}{5} \\ \frac{3}{5} \end{pmatrix} = -\frac{4x_1}{5} \sin\left(\frac{x_1^2}{2}\right) + \frac{6}{5}x_2.$$

Par exemple, dans la direction donnée, cet accroissement ne cesse de changer de signe selon la valeur de x_1 .

Avec ces notions, on peut alors définir l'approximation linéaire d'une fonction f au voisinage d'un point.

Définition 1.7: Approximation affine

Soit D un sous-ensemble de \mathbb{R}^n et f une fonction de D dans \mathbb{R} .

Alors f possède une approximation affine en $\mathbf{a} \in D$, s'il existe une forme linéaire L telle que pour tout vecteur \mathbf{h} vérifiant $\mathbf{a} + \mathbf{h} \in D$ on ait

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + L(\mathbf{h}) + o(\|\mathbf{h}\|),$$

où $o(\|\mathbf{h}\|)$ représente une fonction qui tend vers 0 lorsque \mathbf{h} tend vers $\mathbf{0}$ plus vite que $\|\mathbf{h}\|$.

De plus, si f est de classe C^1 sur D , alors on

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \mathbf{h} \rangle + o(\|\mathbf{h}\|).$$

Si on pose $\mathbf{h} = \mathbf{x} - \mathbf{a}$ dans la définition précédente, nous sommes alors en mesure d'approcher la valeur de la fonction f en n'importe quel point \mathbf{x} à partir de sa valeur en \mathbf{a} de sa dérivée en ce point.

Ainsi, comme dans le cas uni-dimensionnel, l'équation

$$f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \mathbf{x} - \mathbf{a} \rangle$$

définit l'**hyperplan tangent** à f en \mathbf{a} .

Exemple 1.10. Considérons la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x_1, x_2) = x_1^2 + x_2^2$.

Et considérons le vecteur de \mathbb{R}^2 $\mathbf{a} = (1/2, 1/2)$.

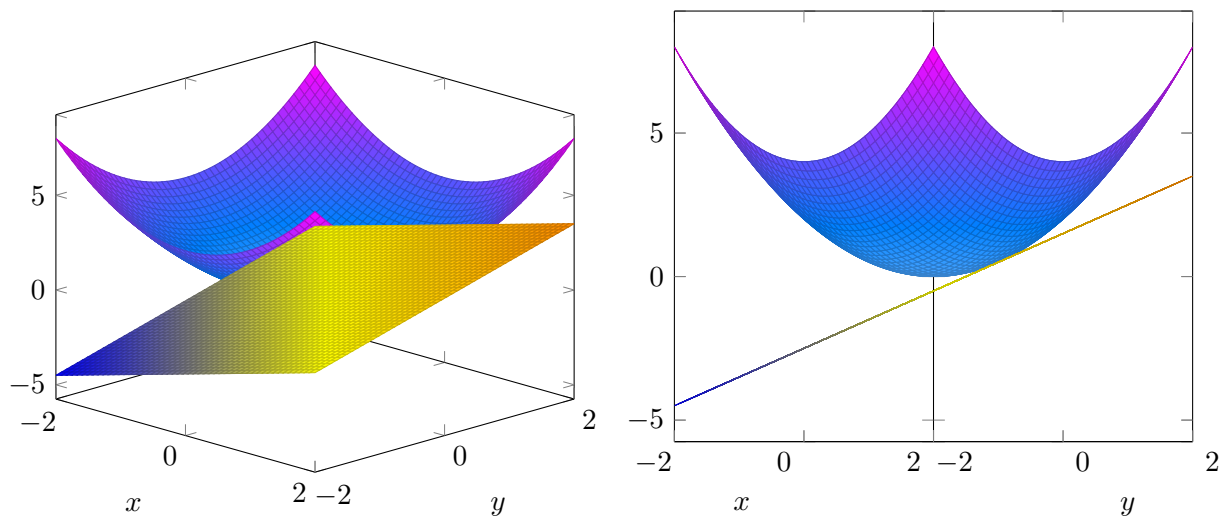
Alors le gradient de la fonction f est donnée par

$$\nabla f(\mathbf{x})^\top = (2x_1 \quad 2x_2).$$

Et le plan tangent de f au point \mathbf{a} est donné par l'équation

$$T(\mathbf{x}) = f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \mathbf{x} - \mathbf{a} \rangle = -\frac{1}{2} + x + y.$$

Ce que l'on peut représenter graphiquement comme suit



Dans la suite, nous allons nous concentrer sur les fonctions qui sont cette fois-ci deux fois différentiables.

1.2 Fonctions différentiables d'ordre 2

Le fait, pour une fonction d'être deux fois différentiable va nous permettre de fournir une caractérisation des fonctions convexes. Nous pourrons aussi de déterminer la nature d'un extremum d'une fonction, comme nous l'avons fait dans le cas des fonctions réelles à valeurs réelles.

Commençons par regarder ce qu'est une fonction deux fois différentiables en définissant les dérivées partielles d'ordre 2.

Définition 1.8: Dérivées partielles d'ordre 2

Soit D un sous-ensemble de \mathbb{R}^n et $\mathbf{x} \in \mathbb{R}^n$. Considérons une fonction f de D dans \mathbb{R} dont les dérivées partielles d'ordre 1 admettent des dérivées partielles, i.e., telle que pour $i \in \llbracket 1, n \rrbracket$, $\frac{\partial f}{\partial x_i}(\mathbf{x})$ existent et soient différentiables.

Cette dérivée est notée $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x})$.

Si les dérivées premières partielles admettent des dérivées partielles d'ordre 1, alors les fonctions

$$\forall i, j \in \llbracket 1, n \rrbracket, \frac{\partial^2 f}{\partial x_j \partial x_i}$$

sont appelées **dérivées partielles d'ordre 2**

Exemple 1.11. Reprenons la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par

$$f(x, y) = xy^2 - yx^2.$$

Cette fonction admet des dérivées partielles d'ordre 2. Les dérivées partielles d'ordre 1 sont

$$\frac{\partial f}{\partial x}(x, y) = y^2 - 2yx \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = 2xy - x^2.$$

Ainsi, les dérivées partielles d'ordre 2 sont définies par, lorsque l'on dérive deux fois par rapport à la même variable

$$\frac{\partial^2 f}{\partial x \partial x}(x, y) = \frac{\partial^2 f}{\partial x^2}(x, y) = -2y \quad \text{et} \quad \frac{\partial^2 f}{\partial y \partial y}(x, y) = \frac{\partial^2 f}{\partial y^2}(x, y) = 2x.$$

Et les dérivées dites croisées sont données par

$$\frac{\partial f}{\partial y} \left(\frac{\partial f}{\partial x} \right) (x, y) = 2y - 2x \quad \text{et} \quad \frac{\partial f}{\partial x} \left(\frac{\partial f}{\partial y} \right) (x, y) = 2y - 2x.$$

Dans cet exemple, on remarque que les dérivées partielles croisées sont égales. Ce n'est pas propre à cet exemple et c'est une conséquence d'un résultat plus général que l'on appelle le **Théorème de Schwarz**.

Théorème 1.1: Théorème de Schwarz

Soit D un sous-ensemble de \mathbb{R}^n et f une application de classe C^2 de D dans \mathbb{R}^n .
Alors

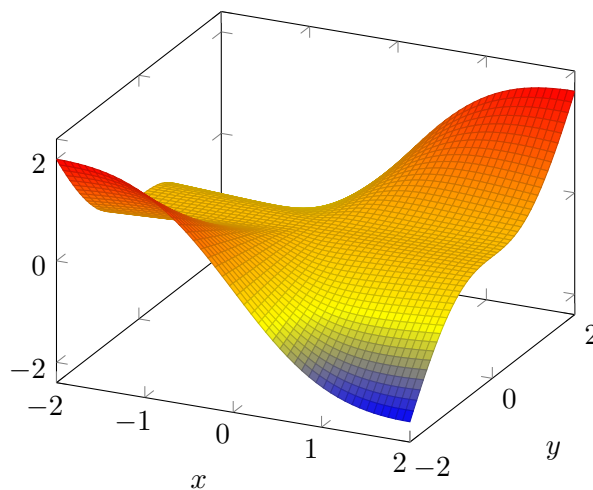
$$\forall i, j \llbracket 1, n \rrbracket \frac{\partial f}{\partial y} \left(\frac{\partial f}{\partial x} \right) = \frac{\partial f}{\partial x} \left(\frac{\partial f}{\partial y} \right).$$

Ce théorème nous dit que dans le cas où la fonction étudiée est de classe C^2 , donc les dérivées partielles d'ordre 2 existent et sont continues, alors l'ordre de dérivation n'a pas d'importance.

Ce résultat peut également servir à montrer qu'une fonction f n'est pas de classe C^2 .

Exemple 1.12. Soit f la fonction définie de \mathbb{R}^2 dans \mathbb{R} par

$$f(x, y) = \frac{xy^3}{x^2 + y^2} \text{ si } (x, y) \neq (0, 0) \text{ et } f(0, 0) = 0$$



La fonction f est bien de classe C^1 . Les dérivées partielles sont nulles en $(0, 0)$ et en dehors, nous avons

$$\frac{\partial f}{\partial x}(x, y) = y^3 \frac{y^2 - x^2}{(x^2 + y^2)^2} \text{ et } \frac{\partial f}{\partial y}(x, y) = xy^2 \frac{3x^2 + y^2}{(x^2 + y^2)^2}.$$

Pour tout y non nul et x non nul respectivement, nous avons

$$\frac{\partial f}{\partial x}(0, y) = -y \quad \text{et} \quad \frac{\partial f}{\partial y}(x, 0) = x.$$

Ainsi

$$\lim_{y \rightarrow 0} \frac{\frac{\partial f}{\partial x}(0, y) - \frac{\partial f}{\partial x}(0, 0)}{y - 0} = -1$$

et

$$\lim_{x \rightarrow 0} \frac{\frac{\partial f}{\partial y}(x, 0) - \frac{\partial f}{\partial y}(0, 0)}{x - 0} = 1$$

Les dérivées partielles secondes en $(0, 0)$ ne sont pas égales. La fonction n'est donc pas de classe C^2 .

Tout comme nous avons défini le Jacobien d'une fonction f , quand cette dernière était de classe C^1 , nous définissons ce que l'on appelle une **matrice hessienne**, pour les fonctions de classe C^2 .

Définition 1.9: Matrice Hessienne

Soit f une fonction définie sur un sous-ensemble D de \mathbb{R}^n à valeurs dans \mathbb{R} et de classe C^2 . On appelle **matrice hessienne** ou parfois **hessien** de f en un point $\mathbf{x} \in \mathbb{R}^n$, la matrice notée

$$\nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \right)_{i,j=1,\dots,n}.$$

On peut également rencontrer la notation $Hess$ pour désigner le hessien de la fonction f .

Plus généralement, si une fonction f est de classe C^k , on peut à nouveau approcher cette fonction par un développement de Taylor d'ordre au plus k .

Définition 1.10: Formules de Taylor-Young

Soit f une fonction définie sur un sous-ensemble D de \mathbb{R}^n à valeurs réelles de classe C^k et \mathbf{x} un point de \mathbb{R}^n ,

Ainsi, une approximation d'ordre 2 d'une fonction f de classe C^2 au voisinage de $\mathbf{x} \in \mathbb{R}^2$ est donnée, $\forall \mathbf{h} = (h_1, h_2) \in \mathbb{R}^2$, par

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + h_1 \frac{\partial f}{\partial x_1}(\mathbf{x}) + h_2 \frac{\partial f}{\partial x_2}(\mathbf{x})$$

$$\begin{aligned}
& + \frac{1}{2} \left(h_1^2 \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) + 2h_1 h_2 \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) \right) + h_2^2 \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) + o(\|\mathbf{h}\|^2), \\
& = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \langle \nabla^2 f(\mathbf{x}) \mathbf{h}, \mathbf{h} \rangle + o(\|\mathbf{h}\|^2).
\end{aligned}$$

On servira de cette approximation lorsque l'on cherchera à étudier le fonctionnement d'algorithmes d'optimisation qui utilise la dérivée seconde de la fonction.

Exemple 1.13. Soit la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par

$$f(x, y) = ye^x + \cos(xy).$$

Cette fonction f est bien de classe C^2 sur \mathbb{R}^2 et admet donc un développement limité d'ordre 2 en $\mathbf{0}$ donnée par

$$f(\mathbf{x}) = f(\mathbf{0}) + x \frac{\partial f}{\partial x}(\mathbf{0}) + y \frac{\partial f}{\partial y}(\mathbf{0}) + \frac{1}{2} \left(x^2 \frac{\partial^2 f}{\partial x^2}(\mathbf{0}) + 2xyx \frac{\partial^2 f}{\partial x \partial y}(\mathbf{0}) + y^2 x \frac{\partial^2 f}{\partial y^2}(\mathbf{0}) \right),$$

où la gradient est donné par

$$\nabla f(\mathbf{x}) = (ye^x - y \sin(xy) \quad e^x - x \sin(xy)) \quad \text{soit} \quad \nabla f(\mathbf{0}) = (0 \quad 1).$$

La matrice hessienne est donnée par

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} ye^x - y^2 \cos(xy) & e^x - \sin(xy) - xy \cos(xy) \\ e^x - \sin(xy) - xy \cos(xy) & -x^2 \cos(xy) \end{pmatrix} \quad \text{soit} \quad \nabla^2 f(\mathbf{0}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Donc une approximation de f d'ordre 2 au point $\mathbf{0}$ est donnée par

$$f(\mathbf{x}) = 1 + y + 2xy.$$

Nous avons maintenant tous les outils nécessaires à la caractérisations des extrema d'une fonction, ainsi qu'à l'étude des fonctions convexes.

2 Recherche d'extrema et applications aux fonctions convexes

On souhaite maintenant regarder comment les outils de calculs différentielles, *i.e.*, l'étude de la dérivée des fonctions peut nous permettre de localiser et caractériser les extrema d'une fonction.

Nous reviendrons également sur les notions de convexité et la caractérisation de la convexité.

2.1 Retour sur la convexité

Commençons par un retour sur la notion de convexité, elle sera essentielle dans la caractérisation d'un extremum d'une fonction.

Définition 2.1: Convexité

Soit f une fonction définie sur un domaine D de \mathbb{R}^d à valeurs dans \mathbb{R} . On dit que la fonction f est **convexe** sur D si pour tout $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ et pour tout $\alpha \in [0, 1]$ nous avons

$$f(\alpha \mathbf{u} + (1 - \alpha) \mathbf{v}) \leq \alpha f(\mathbf{u}) + (1 - \alpha) f(\mathbf{v}).$$

Dans cette définition, si on remplace le signe \leq par \geq , on définit une fonction **concave**. De plus, si les inégalités larges sont remplacées par des inégalités strictes, la fonction sera dite **strictement convexe** ou **strictement concave**. On rappelle également que si une fonction f est convexe alors $-f$ est concave et réciproquement.

On peut également caractériser la convexité d'une fonction en se basant sur le gradient de cette dernière ou encore en utilisant la matrice hessienne.

Définition 2.2: Convexité, caractérisation ordre 1

Soit f une fonction de classe C^1 sur un ensemble $D \subset \mathbb{R}^d$, alors

- f est **convexe** si tout hyperplan tangent à f est en **dessous** de son graphe, *i.e.*,

$$\forall \mathbf{u}, \mathbf{v} \in D, f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \mathbf{u} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle.$$

- f est **concave** si tout hyperplan tangent à f est au **dessus** de son graphe, *i.e.*,

$$\forall \mathbf{u}, \mathbf{v} \in D, f(\mathbf{u}) \leq f(\mathbf{v}) + \langle \mathbf{u} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle.$$

Enfin, la caractérisation d'ordre 2.

Définition 2.3: Convexité, caractérisation ordre 2

Soit f une fonction de classe C^2 sur un ensemble $D \subset \mathbb{R}^d$, alors

- f est **convexe** si sa matrice hessienne est *semie-définie positive*, i.e.,

$$\forall \mathbf{u}, \mathbf{v} \in D, \mathbf{u}^\top \nabla^2 f(\mathbf{v}) \mathbf{u} \geq 0.$$

- f est **concave** si sa matrice hessienne est *semie-définie négative*, i.e.,

$$\forall \mathbf{u}, \mathbf{v} \in D, \mathbf{u}^\top \nabla^2 f(\mathbf{v}) \mathbf{u} \leq 0.$$

Exemple 2.1. Soit f la fonction définie de \mathbb{R}^2 dans \mathbb{R} définie par

$$f(x, y) = (4 - 2y)^2 + 5x^2 + x + 3y + 4xy.$$

On va étudier la convexité de cette dernière.

Commençons par voir que la fonction f est une forme quadratique que l'on peut donc réécrire sous forme matricielle.

$$\begin{aligned} f(x, y) &= (4 - 2y)^2 + 5x^2 + x + 3y + 4xy, \\ &= 16 - 16y + 4y^2 + 5x^2 + x + 3y + 4xy, \\ &= 5x^2 + 4y^2 + 4xy + x - 13y + 16, \\ &= \langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{b}, \mathbf{u} \rangle + 16, \end{aligned}$$

$$\text{où } \mathbf{u} = \begin{pmatrix} x & y \end{pmatrix}^\top, \mathbf{A} = \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix} \text{ et } \mathbf{b} = \begin{pmatrix} 1 & -13 \end{pmatrix}^\top$$

Pour cette forme quadratique, la matrice hessienne est donnée par $2\mathbf{A}$, or cette matrice \mathbf{A} a une trace et un déterminant positif, ses valeurs propres sont donc positives et la matrice \mathbf{A} est donc positive. Ainsi f est convexe.

2.2 Conditions d'optimalité

Après ce bref rappel sur les fonctions convexes, nous lançons dans la recherche d'extrema d'une fonction f définie de $D \subset \mathbb{R}^d$ à valeurs dans \mathbb{R} .

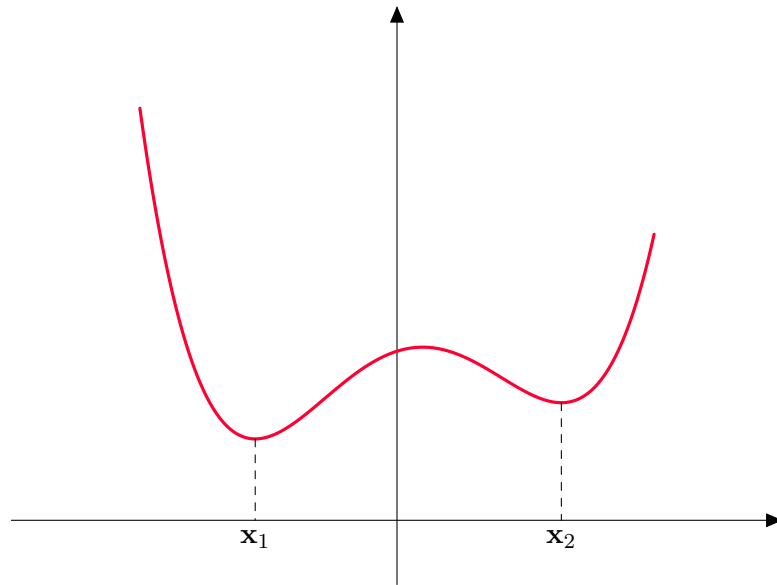


FIGURE 1 – Illustration du concept de minimum local et global d’une fonction. Le point \mathbf{x}_1 est le minimum global de la fonction alors que \mathbf{x}_2 est un simple minimum local.

Définition 2.4: Minimum local et global

Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction continue. On dit que $\mathbf{u} \in \mathbb{R}^d$ est **minimum local** de f si pour tout voisinage $V \subset \mathbb{R}^d$ de \mathbf{u} (c’est une ensemble qui contient de le point \mathbf{u}) on a :

$$f(\mathbf{u}) \leq f(\mathbf{v}), \quad \forall \mathbf{v} \in V,$$

Le point \mathbf{u} est un **minimum global** de la fonction f si et seulement si :

$$f(\mathbf{u}) \leq f(\mathbf{v}), \quad \forall \mathbf{v} \in \mathbb{R}^d.$$

Noter bien la différence entre local et global, dans le premier cas, on supposera que la fonction a atteint son minimum dans un espace donné. Dans le second cas, elle ne doit pas atteindre une valeur plus faible que celle atteinte en \mathbf{u} .

En d’autres termes, le minimum d’une fonction f est juste la valeur $f(\mathbf{u})$ où \mathbf{u} est le point en lequel la fonction f atteint son minimum (local ou global) comme illustré par la Figure 1.

Lorsque l’on peut représenter le graphe de la fonction, il est très simple de localiser un minimum (ou un maximum). Mais cela devient plus difficile en dimension supérieure.

La proposition suivante permet de donner une première caractérisation d'un minimum (local) d'une fonction f . C'est ce que l'on appelle l'**inéquation d'Euler**.

Proposition 2.1: Inéquation d'Euler

Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction continue et \mathcal{U} un ensemble convexe non vide de \mathbb{R}^d . Considérons également $\mathbf{u} \in \mathcal{U}$ un minimum de relatif de f par rapport à \mathcal{U} . Si f est différentiable en \mathbf{u} , alors

$$\nabla \langle f(\mathbf{u}), (\mathbf{v} - \mathbf{u}) \rangle \geq 0, \forall \mathbf{v} \in \mathcal{U}$$

Cela signifie simplement que si on évalue le gradient de f en \mathbf{u} alors la fonction sera croissante dans n'importe quelle direction $\mathbf{v} - \mathbf{u}$

Démonstration. Supposons que \mathbf{u} est un minimiseur relatif par rapport à notre ensemble \mathcal{U} . Alors pour tout $\alpha > 0$ et tout vecteur \mathbf{v} tels que $\alpha(\mathbf{u} - \mathbf{v}) \in \mathcal{U}$, nous avons

$$f(\mathbf{u}) \leq f(\mathbf{u} + \alpha(\mathbf{u} - \mathbf{v})).$$

Ainsi

$$\frac{f(\mathbf{u}) - f(\mathbf{u} + \alpha(\mathbf{u} - \mathbf{v}))}{\alpha} \geq 0.$$

En prenant la limite lorsque α tend vers 0, on trouve que

$$D_{\mathbf{u}} f(\mathbf{u}) \geq 0.$$

□

Mais ce résultat est rarement utilisé en pratique Si on revient à la Figure 1, on remarque que les extrema sont localisés en les points où le gradient de la fonction f s'annule. Cette condition d'optimalité est connue sous le d'**équation d'Euler**.

Proposition 2.2: Equation d'Euler

Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction continue et différentiable en $\mathbf{u} \in \mathbb{R}^d$. Si \mathbf{u} est un extremum alors :

$$\nabla f(\mathbf{u}) = 0.$$

Démonstration. On fait la preuve dans le cas où \mathbf{u} est un minimum, mais elle est analogue dans le cas où \mathbf{u} est un maximum.

En utilisant la définition de minimum dans un voisinage de \mathbf{u} , i.e., $\forall \mathbf{v} \in \mathbb{R}^d, \exists t > 0$ tels que $\mathbf{u} + t\mathbf{v} \in V$, nous avons

$$\begin{aligned} f(\mathbf{u}) &\leq f(\mathbf{u} + t\mathbf{v}) = f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (t\mathbf{v}) + t\mathbf{v}^\top \varepsilon(t\mathbf{v}), \quad t \ll 1 \\ \iff 0 &\leq \nabla f(\mathbf{u})^\top (t\mathbf{v}) + t\mathbf{v}^\top \varepsilon(t\mathbf{v}), \end{aligned}$$

où ε est une fonction vectorielle dont les composantes tendent vers 0 lorsque t tend vers 0.

En divisant par $t > 0$ et en prenant la limite lorsque $t \rightarrow 0$, on obtient :

$$0 \leq \nabla f(\mathbf{u})^\top \mathbf{v}$$

En remplaçant maintenant \mathbf{v} par $-\mathbf{v}$ on a, de façon similaire :

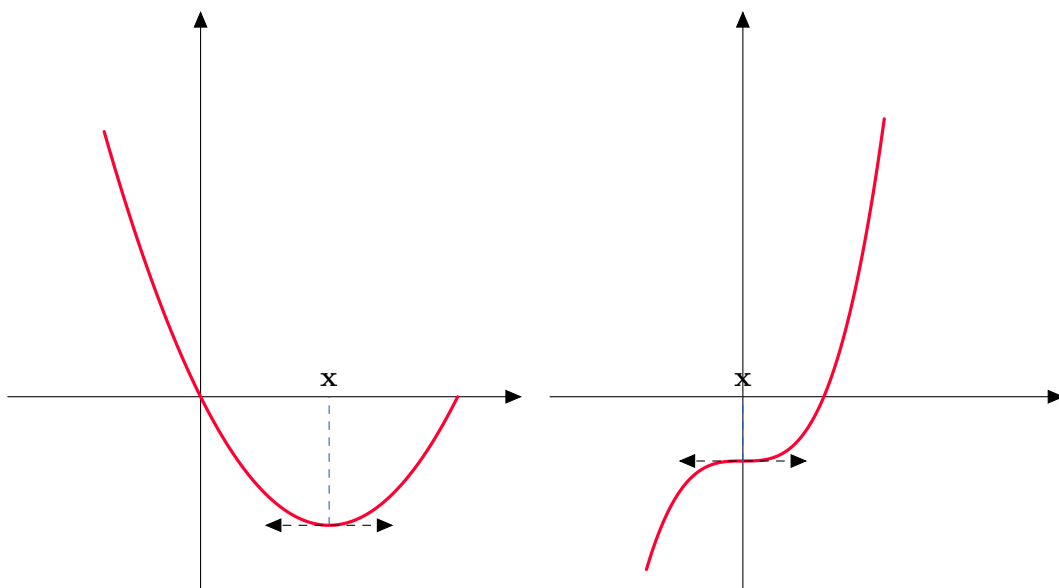
$$0 \leq -\nabla f(\mathbf{u})^\top \mathbf{v}.$$

Ainsi pour tout $\mathbf{v} \in \mathbb{R}^d$ on a $\nabla f(\mathbf{u})^\top \mathbf{v} = 0$ et donc $\nabla f(\mathbf{u}) = 0$. □

Les solutions de l'équation d'Euler définissent ce que l'on appelle des *points critiques* ou des *valeurs critiques*.

Faites bien attention ! Cette proposition ne donne qu'une condition nécessaire pour trouver un extremum d'une fonction, mais un point critique n'est pas forcément un extremum, la condition n'est pas suffisante comme nous avons pu le voir dans le cas unidimensionnel.

Exemple 2.2. Soient les fonctions f et g respectivement définies par $\frac{1}{2}(x-2)^2 - 2$ and $\frac{1}{2}x^3 - 1$ dont la représentation est donnée ci-dessous.



Les points \mathbf{x} des deux graphes sont solutions de l'équation d'Euler $\nabla f(\mathbf{x}) = 0$ et $\nabla g(\mathbf{x}) = 0$ respectivement. Cependant, s'il s'agit bien d'un point en lequel la fonction f atteint son minimum, on ne peut pas dire que le point \mathbf{x} soit un point en lequel la fonction g atteint un extremum.

Cet exemple montre qu'il est important de définir des critères qui permettent de caractériser la nature d'un point critique : minimum, maximum ou ... ni l'un ni l'autre.

Caractérisation des extrema

Etant donnée une solution \mathbf{u} de $\nabla f(\mathbf{u}) = 0$, on peut dire que :

- \mathbf{u} est un **minimum local** si $\nabla^2 f(\mathbf{u}) = \text{Hess}_f(\mathbf{u}) \geq 0$, *i.e.* si la matrice hessienne de la fonction f évaluée en \mathbf{u} est semie-définie positive. Cela signifie qu'en ce point, la fonction f est localement convexe!
Ce point est un minimum global si f est **convexe** sur son ensemble de définition ou si pour tout $\mathbf{v} \neq \mathbf{u}$ on a $f(\mathbf{u}) \leq f(\mathbf{v})$.
- \mathbf{u} est un **maximum local** si $\nabla^2 f(\mathbf{u}) = \text{Hess}_f(\mathbf{u}) \leq 0$, *i.e.* si la matrice hessienne de la fonction f évaluée en \mathbf{u} est semie-définie négative. Cela signifie qu'en ce point, la fonction f est localement concave!
Ce point est un maximum global si f est **concave** sur son ensemble de définition ou si pour tout $\mathbf{v} \neq \mathbf{u}$ on a $f(\mathbf{u}) \geq f(\mathbf{v})$.
- Dans les autres cas, on ne peut rien dire ou il faut procéder à une étude

plus poussée.

Exemple 2.3. Soit f la fonction définie de \mathbb{R}^2 dans \mathbb{R} définie par

$$f(x, y) = (4 - 2y)^2 + 5x^2 + x + 3y + 4xy.$$

On va étudier la convexité de la fonction f et chercher ses extremum.

Commençons par voir que la fonction f est une forme quadratique que l'on peut donc réécrire sous forme matricielle.

$$\begin{aligned} f(x, y) &= (4 - 2y)^2 + 5x^2 + x + 3y + 4xy, \\ &= 16 - 16y + 4y^2 + 5x^2 + x + 3y + 4xy, \\ &= 5x^2 + 4y^2 + 4xy + x - 13y + 16, \\ &= \langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{b}, \mathbf{u} \rangle + 16, \end{aligned}$$

$$\text{où } \mathbf{u} = \begin{pmatrix} x & y \end{pmatrix}^\top, \mathbf{A} = \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix} \text{ et } \mathbf{b} = \begin{pmatrix} 1 & -13 \end{pmatrix}^\top$$

Pour cette forme quadratique, la matrice hessienne est donnée par $2\mathbf{A}$, or cette matrice \mathbf{A} a une trace et un déterminant positif, ses valeurs propres sont donc positives et la matrice A est donc positive. Ainsi f est convexe.

Pour rechercher ses extrema, on commence par résoudre l'équation d'Euler

$$\nabla f(\mathbf{u}) = 2\mathbf{A}\mathbf{u} + \mathbf{b} = \mathbf{0}.$$

Cela nous amène à résoudre le système

$$\begin{cases} 10x + 4y + 1 &= 0, \\ 4x + 8y - 13 &= 0. \end{cases}$$

La solution de ce système est donnée par le vecteur

$$\mathbf{u} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -\frac{30}{32} \\ \frac{67}{32} \end{pmatrix}.$$

La fonction f étant convexe, il s'agit du minimum global de la fonction f .

Etudions un deuxième exemple qui n'implique pas une forme quadratique.

Exemple 2.4. Soit f la fonction définie de \mathbb{R}^2 dans \mathbb{R} définie par

$$f(\mathbf{u}) = f(x, y) = 2x^2 + 4(y - 2)^2 + 4x + 6y - 2xy + 2y^3.$$

On va reprendre la même étude que précédemment.

Nous avons un terme cubique donc on ne peut pas exprimer cette fonction à l'aide d'une matrice comme nous l'avons fait dans l'exemple précédent.

La matrice hessienne de notre fonction est donnée par

$$\text{Hess}_f(\mathbf{u}) = \begin{pmatrix} 4 & -2 \\ -2 & 12y + 8 \end{pmatrix}$$

Remarquons que la trace et le déterminant de cette matrice sont respectivement égales à $12y + 12$ et $48y + 28$. Ces deux quantités sont positives si et seulement si $y \geq -\frac{7}{12}$.

Ainsi la fonction est convexe sur $\mathbb{R} \times \left[-\frac{7}{12}, +\infty\right[$.

Elle est concave si $12y + 12$ est négatif et $48y + 28$ est positif, donc si y vérifie $y \leq -1$ et $y \geq -\frac{7}{12}$. Ces deux conditions sont incompatibles donc la fonction n'est ni convexe, ni concave en dehors de l'ensemble $\mathbb{R} \times \left[-\frac{7}{12}, +\infty\right[$.

Pour trouver les extrema de cette fonction, on doit résoudre le système

$$\nabla f(\mathbf{u}) = 0 \iff \begin{cases} 4x + 4 - 2y & = & 0, \\ 6y^2 + 8y - 2x - 10 & = & 0. \end{cases} \iff \begin{cases} 2x & = & y - 2, \\ 6y^2 + 7y - 8 & = & 0. \end{cases}$$

Ce système admet deux solutions \mathbf{u}_1 et \mathbf{u}_2 qui sont

$$\mathbf{u}_1 = \begin{pmatrix} \frac{-31}{24} - \frac{\sqrt{241}}{7} \\ -\frac{7}{12} - \frac{\sqrt{241}}{12} \end{pmatrix} \quad \text{et} \quad \mathbf{u}_2 = \begin{pmatrix} \frac{-31}{24} + \frac{\sqrt{241}}{7} \\ -\frac{7}{12} + \frac{\sqrt{241}}{12} \end{pmatrix}$$

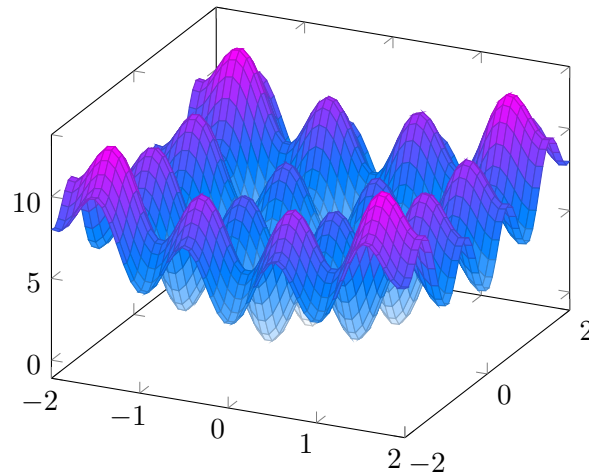
Le premier point \mathbf{u}_1 voit sa deuxième composante être plus petite que -1 , il ne s'agit donc ni d'un minimum, ni d'un maximum. Quant au deuxième point, sa deuxième composante est strictement positive, il s'agit donc d'un **minimum local** de cette fonction. Ce n'est pas un minimum global, car la fonction f tend vers $-\infty$ lorsque y tend vers $-\infty$.

Regardons un dernier exemple

Exemple 2.5. Etudions $f : [-2, 2]^2 \rightarrow \mathbb{R}$ définie par

$$f(\mathbf{x}) = 4 + (x_1^2 - 2 \cos(2\pi x_1)) + (x_2^2 - 2 \cos(2\pi x_2)).$$

Cette fonction est intéressante car elle admet plusieurs minima locaux mais un seul minimum global. Elle est souvent utilisée pour tester l'efficacité d'un algorithme d'optimisation.



On peut voir que cette fonction n'est ni convexe, ni concave globalement. On peut regarder le gradient de cette fonction

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 + 4\pi \sin(2\pi x_1) & 2x_2 + 4\pi \sin(2\pi x_2) \end{pmatrix}.$$

Les solutions de l'équation d'Euler sont données par la résolution du système

$$\begin{aligned} x_1 &= -2\pi \sin(2\pi x_1), \\ x_2 &= -2\pi \sin(2\pi x_2). \end{aligned}$$

Les deux équations sont indépendantes mais sont globalement difficiles à résoudre ... sauf si on garde l'esprit que $\sin(x) \in [-1, 1]$. Ainsi, on peut voir que le vecteur nul est un point critique, qui est aussi le minimum global de la fonction.

On va maintenant regarder quelques exemples concrets de problèmes que l'on va chercher à résoudre en Science des Données.

Deuxième partie

Modèles Linéaires Gaussiens

Nous plaçons à présent dans un cadre bien précis, où nous cherchons à expliquer les valeurs prises par une variable aléatoire quantitative Y en fonction des valeurs prises par un ensemble de variables X_1, X_2, \dots, X_p quantitatives ou qualitatives telles que

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

où $\mathbf{X} = (X_1, X_2, \dots, X_p)$ et où, d'où le nom **Gaussien**, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ où σ^2 est inconnue.

Nous ferons également l'hypothèse que la fonction f considérée est une fonction **linéaire**, *i.e.*, notre modèle peut s'écrire

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon.$$

Les valeurs du vecteur $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ sont alors à déterminer.

Pour cela nous disposons d'un jeu de données qui va nous permettre d'obtenir une estimation de ces paramètres à l'aide d'un critère que l'on va définir et que l'on cherchera à minimiser. Cela se présentera sous la forme d'un problème d'optimisation

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \varphi(\boldsymbol{\beta}).$$

On cherchera à résoudre ce problème avec un échantillon de données $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ et $\mathbf{x}_i \in \mathbb{R}^p$ tel que

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \dots + \beta_p X_{1,p} + \varepsilon_1, \\ y_2 &= \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \dots + \beta_p X_{2,p} + \varepsilon_2, \\ &\dots = \dots \\ y_{n-1} &= \beta_0 + \beta_1 X_{n-1,1} + \beta_2 X_{n-1,2} + \dots + \beta_p X_{n-1,p} + \varepsilon_{n-1}, \\ y_n &= \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \dots + \beta_p X_{n,p} + \varepsilon_n, \end{aligned}$$

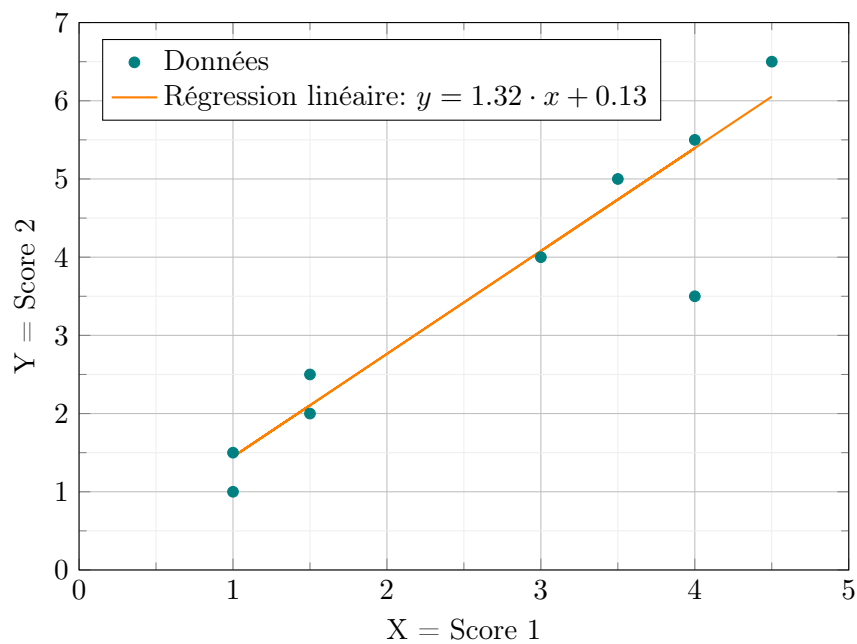
où $X_{i,j}$ indique la j -ème caractéristique de l'individu i .

Regardons un exemple où l'on souhaite prédire le score obtenu à un deuxième examen, en fonction du score obtenu à un premier examen pour un ensemble de 10 étudiants.

Nos données se présentent comme suit

Y : Score examen 2	3.5	4	5	1	2	1.5	2.5	5.5	6	6.5
X : Score examen 1	4	3	3.5	1	1.5	1	1.5	4	3.5	4.5

L'objectif sera ici d'apprendre les coefficients de la droite de régression. Nous pouvons représenter graphiquement les données sur le graphe ci-dessous, ainsi que la droite obtenue



L'objectif sera d'étudier comment nous pouvons déterminer ces coefficients à l'aide notre jeu de données.

Nous pousserons également notre étude en étudiant les propriétés statistiques des estimateurs $\hat{\beta}$ obtenus de β , construire des intervalles de confiance sur ces derniers ou encore sur les prédictions effectuées par le modèle.

3 Modèle Linéaire Gaussien simple

Dans cette première partie, on va s'intéresser au modèle linéaire **simple**, c'est-à-dire le modèle où l'on cherche à prédire les valeurs de la variable Y uniquement en fonction d'une seule variable X .

Notre modèle s'écrira donc

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où β_0 représente l'ordonnée à l'origine du modèle (*intercept* en anglais) et β_1 représente le coefficient directeur de notre droite (*slope* en anglais). C'est ce coefficient β_1 qui va décrire l'impact de la variable X sur la variable Y .

Pour être plus exact d'un point de vue mathématiques, nous devrions employer le terme *affine* et non *linéaire*, dans la mesure où la droite apprise ne passe pas forcément par l'origine, sauf si $\beta_0 = 0$.

Dans la suite, nous supposons également que la variable explicative X suit une distribution normale. Essayons maintenant de comprendre comment nous pouvons estimer ces paramètres.

3.1 Hypothèse du modèle Gaussien

Notre modèle gaussien *simple* (car on n'utilise qu'une seule variable) va chercher expliquer la relation qui existe entre deux variables quantitatives X et Y par une relation affine

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où $Y \in \mathbb{R}$ désigne la variable à expliquer, $X \in \mathbb{R}$ la variable explicative, $\beta = (\beta_0, \beta_1)$ les paramètres du modèle que l'on cherche à estimer et ε est un terme d'erreur aléatoire.

Nous avons dit que pour estimer les paramètres du modèle, nous utilisons un jeu de données $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, i.e.,

$$\forall i \in \llbracket 1, n \rrbracket, Y_i = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i.$$

Hypothèses Modèle Gaussien

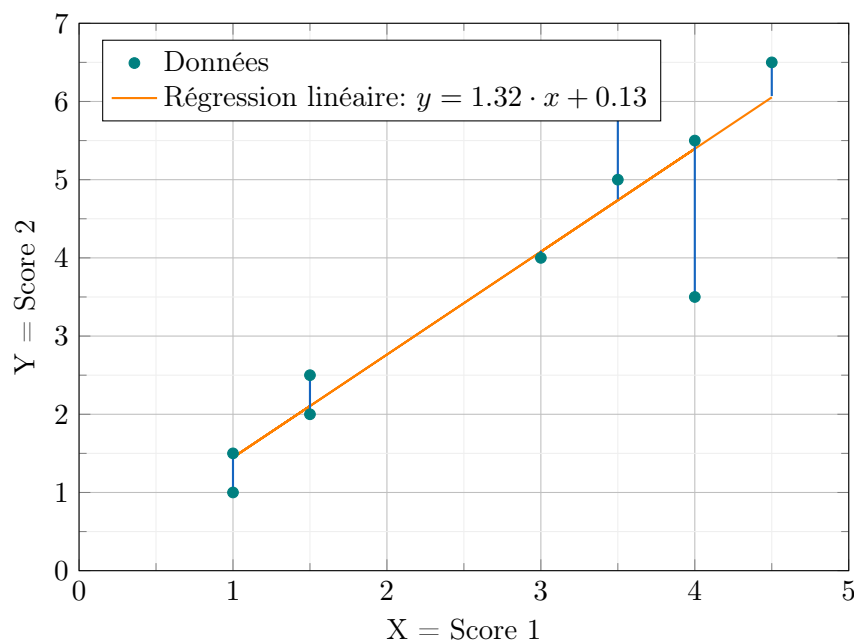
Nous formulons les hypothèses suivantes pour notre modèle linéaire gaussien :

1. $(Y_i, X_i)_{i=1}^n$ doivent être *i.i.d.*, *i.e.* indépendantes et identiquement distribuées,
2. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$
3. $\varepsilon_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$: hypothèse d'homoscédasticité.

Les première et deuxième hypothèses précisent que les valeurs Y_i sont **observées et aléatoires** alors que les valeurs X_i sont **observées et non aléatoires** (on dit aussi déterministes). La troisième hypothèse précise que les erreurs sont aléatoires et distribuées selon une loi gaussienne : (i) centrée, (ii) de variance inconnue σ^2 et (iii) indépendantes. Ce dernier point signifie que la **covariance** entre les erreurs associées à des individus i et j est nulle, *i.e.*, $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

3.2 Optimisation

Notre objectif est de déterminer les valeurs des paramètres du modèle telles que la valeur prédite $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ soit la plus proche de la valeur y_i pour les différents individus $x_i, i = 1, \dots, n$.



Nous pourrions donc être tentés d'évaluer cet écart $y_i - \hat{y}_i = \varepsilon_i$ sur l'ensemble des individus, *i.e.*, nous pourrions chercher à résoudre le problème

$$\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n \varepsilon_i = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n y_i - \hat{y}_i = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n y_i - (\beta_0 + \beta_1 x_i).$$

Mais cela ne serait pas une bonne définition des erreurs des modèles, que l'on appelle aussi **les résidus**. En effet, les erreurs devraient être comptés positivement or ici $\varepsilon_i = y_i - \hat{y}_i$ peuvent être positives ou négatives, ce qui entraîne des phénomènes de compensations. De plus, par définition, ces erreurs sont centrées, donc la somme des erreurs, ainsi définies, serait égale à 0.

Nous pourrions prendre la **valeur absolue de la différence** entre l'observation y_i et la prédiction \hat{y}_i , *i.e.*

$$\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n |y_i - (\beta_0 + \beta_1 x_i)|,$$

mais ce problème reste difficile à résoudre sur le plan mathématiques. Nous préférons donc minimiser le **carré de la distance** entre y_i et \hat{y}_i . On parle de recherche des paramètres par la **méthodes moindres carrés**. Cette méthode a un avantage considérable, contrairement à la méthode par maximum de vraisemblance que l'on verra plus tard, car elle ne nécessite aucune hypothèse sur la distribution des erreurs. On va donc résoudre le problème

$$\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n \varepsilon_i^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

3.3 Expression des solutions

Avant de présenter l'expression des solutions de notre problème d'optimisation, nous rappelons le résultat suivant de probabilités

Rappels de probabilités. On rappelle les résultats suivants de probabilités.

Lemme 3.1: Variance de variables aléatoires

Considérons X et Y des variables aléatoires qui admettent des moments d'ordre 2, *i.e.* qui admettent une variance. Alors,

- (i) concernant la variance d'une variable aléatoire, nous avons la **Formule de Koenig-Huygens** :

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- (ii) la covariance entre deux variables aléatoires est aussi égale à

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y].$$

Démonstration. Nous démontrons les deux points séparément

- (i) **Formule de Koenig-Huygens.** On repart de la définition de la variance

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &\quad \downarrow \text{en développant l'expression} \\ &= \mathbb{E}[X^2 - 2X \mathbb{E}[X] - \mathbb{E}[X]^2], \\ &\quad \downarrow \text{par linéarité de l'espérance} \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2, \\ \text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

- (ii) **Egalité sur la covariance.** Même principe, on repart de la définition de la covariance.

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])], \\ &\quad \downarrow \text{on développe} \\ &= \mathbb{E}[XY - \mathbb{E}[X]Y - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]], \\ &\quad \downarrow \text{linéarité de l'espérance} \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y], \\ \text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

□

Le lemme précédent va nous permettre de fournir une preuve plus simple quant à

l'expression des paramètres optimaux de notre modèle de régression.

Proposition 3.1: Problème de régression linéaire

On considère le problème de régression linéaire gaussien de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Les paramètres a et b sont solutions du problème d'optimisation

$$\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n \varepsilon_i^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Dont les solutions sont donnés par

$$\hat{\beta}_1 = \frac{Cov[X, Y]}{Var[X]} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \hat{\beta}_0 = \mathbb{E}[Y] - \mathbb{E}[X] \times \hat{\beta}_1 = \bar{y} - \hat{\beta}_1 \times \bar{x}.$$

Démonstration. La fonction L que l'on cherche à optimiser est une fonction convexe en les variables β_0 et β_1 , elle admet donc une unique solution. Cette solution est obtenue en résolvant l'équation d'Euler se présentant sous la forme d'un système linéaire

$$\frac{\partial L}{\partial \beta_1} = 0 \iff -2 \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0) x_i = 0, \quad (1)$$

$$\frac{\partial L}{\partial \beta_0} = 0 \iff -2 \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0) = 0. \quad (2)$$

On se concentre sur l'équation (2) pour le moment, on va la développer. Ce qui nous permet d'écrire

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0) &= 0, \\ \iff \sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i - n\beta_0 &= 0, \\ \iff \frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i &= \beta_0, \\ \iff \bar{y} - \beta_1 \bar{x} &= \beta_0. \end{aligned}$$

On obtient une expression de l'estimateur $\hat{\beta}_0$ de β_0 , elle dépend de β_1 dont on va pouvoir déterminer l'expression en injectant la valeur de β dans l'équation (1).

$$\begin{aligned}
& -2 \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0) x_i &= 0, \\
\iff & \sum_{i=1}^n y_i x_i - \beta_1 \sum_{i=1}^n x_i^2 - \beta_0 \sum_{i=1}^n x_i &= 0, \\
\iff & \sum_{i=1}^n y_i x_i - \beta_1 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \left(\frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i \right) &= 0, \\
\iff & \sum_{i=1}^n y_i x_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) - \beta_1 \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) &= 0, \\
\iff & \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2} &= \beta_1, \\
\iff & \frac{\text{Cov}[x, y]}{\text{Var}[x]} &= \beta_1,
\end{aligned}$$

où la dernière équivalence est une conséquence du lemme 3.1

□

Étudions maintenant les propriétés de ces estimateurs.

3.4 Propriétés des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$

Nous cherchons maintenant à étudier le biais des estimateurs obtenus par méthodes des moindres carrés ainsi que leurs variances. Cela nous permettra de construire des statistiques de test qui prouveront, ou non, la significativité des paramètres.

Propriétés de la pente du modèle de régression Commençons d'abord par montrer que $\hat{\beta}_0$ est un estimateur sans biais de a , c'est-à-dire que $\mathbb{E}[\hat{\beta}_0] = \beta_0$. On utilisera le fait que

- $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$
- $\mathbb{E}[\bar{y}] = \beta_0 + \beta_1 \bar{x}$

Ainsi, l'espérance du

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1] &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[y_i - \bar{y}]}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x}) (\mathbb{E}[y_i] - \mathbb{E}[\bar{y}])}{\sum_{i=1}^n (x_i - \bar{x})^2},
\end{aligned}$$

$$\begin{aligned}
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\
&= \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\
\mathbb{E}[\hat{\beta}_1] &= \beta_1.
\end{aligned}$$

On peut maintenant faire de même avec la variance de l'estimateur afin de déterminer son écart-type, ce qui nous servira à tester la significativité de la pente, mais aussi à construire l'intervalle de confiance sur l'estimation du paramètre.

Pour cela on utilisera le fait que l'on peut écrire :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \sum_{i=1}^n \omega_i \varepsilon_i, \quad \text{où} \quad \omega_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Cette égalité est une conséquence du fait que $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ et $\bar{y} = \beta_0 + \beta_1 \bar{x}$.

On peut alors déterminer la variance de notre estimateur à l'aide de cette relation :

$$\begin{aligned}
\text{Var}[\hat{\beta}_1] &= \mathbb{E} \left[(\hat{\beta}_1 - \mathbb{E}[\hat{\beta}_1])^2 \right], \\
&= \mathbb{E} \left[\left(\beta_1 + \sum_{i=1}^n \omega_i \varepsilon_i - \beta_1 \right)^2 \right], \\
&= \mathbb{E} \left[\left(\sum_{i=1}^n \omega_i \varepsilon_i \right)^2 \right], \\
&= \mathbb{E} \left[\sum_{i=1}^n (\omega_i \varepsilon_i)^2 + 2 \sum_{i < i'}^n \varepsilon_i \varepsilon_{i'} \omega_i \omega_{i'} \right], \\
&= \sum_{i=1}^n \underbrace{\mathbb{E}[\varepsilon_i^2]}_{=\text{Var}[\varepsilon_i]=\sigma^2} \omega_i^2 + 2 \sum_{i < i'}^n \underbrace{\mathbb{E}[\varepsilon_i \varepsilon_{i'}]}_{=0} \omega_i \omega_{i'}, \\
\text{Var}[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

La variance de l'estimateur $\text{Var}[\hat{\beta}_1]$ dépend de la variance des erreurs σ^2 qui est pour le moment inconnue. Nous formulerons la même remarque au sujet de la variance de l'estimateur $\text{Var}[\hat{\beta}_0]$ par la suite.

Propriétés de l'ordonnée à l'origine du modèle de régression Nous reprenons la même étude. On peut commencer par remarquer que

$$\hat{\beta}_0 = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}.$$

En effet, repartons de l'expression de l'estimateur

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x}, \\ &\downarrow \text{ on utilise le fait que } \bar{y} = \beta_1\bar{x} + \beta_0 \\ &= \beta_1\bar{x} + \beta_0 - \hat{\beta}_1\bar{x}, \\ &= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}.\end{aligned}$$

Commençons par l'espérance :

$$\begin{aligned}\mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}], \\ &\downarrow \text{ linéarité de l'espérance} \\ &= \mathbb{E}[\beta_0] + \mathbb{E}[(\beta_1 - \hat{\beta}_1)\bar{x}], \\ &\downarrow \text{ seul } \hat{\beta}_1 \text{ est aléatoire ici} \\ &= \beta_0 + (\beta_1 - \mathbb{E}[\hat{\beta}_1])\bar{x}, \\ &\downarrow \text{ on a vu que } \mathbb{E}[\hat{\beta}_1] = \beta_1 \\ &= \beta_0.\end{aligned}$$

Regardons enfin l'espérance de cette estimateur, pour cela on va repartir de la définition de départ

$$\begin{aligned}\text{Var}[\hat{\beta}_0] &= \text{Var} \left[\bar{y} - \hat{\beta}_1\bar{x} \right], \\ &\downarrow \text{ définition de la variance} \\ &= \text{Var} \left[\frac{1}{n} \sum_{i=1}^n y_i \right] + \text{Var} [\hat{\beta}_1\bar{x}] - 2\text{Cov} \left(\frac{\sum_{i=1}^n y_i}{n}, \hat{\beta}_1\bar{x} \right), \\ &\downarrow \text{ seul les } y_i \text{ et } \hat{\beta}_1 \text{ sont aléatoires} \\ &= \frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n y_i \right] + \bar{x}^2 \text{Var}[\hat{\beta}_1] - \frac{2\bar{x}}{n} \text{Cov} \left(\sum_{i=1}^n y_i, \hat{\beta}_1 \right), \\ &\downarrow \text{ or } \text{Var}[y_i] = \sigma^2 \text{ et les } y_i \text{ sont indépendants} \\ &\downarrow \text{ on a déjà calculé la variance de } \hat{\beta}_1\end{aligned}$$

$$\begin{aligned}
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2\bar{x}}{n} Cov \left(\sum_{i=1}^n y_i, \sum_{i=1}^n \omega_i \varepsilon_i \right), \\
&\quad \downarrow \text{ on utilise le fait } y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2\bar{x}}{n} Cov \left(\sum_{i=1}^n \varepsilon_i, \sum_{i=1}^n \omega_i \varepsilon_i \right), \\
&\quad \downarrow \text{ on utilise le fait que les erreurs sont indépendantes, i.e. } Cov(\varepsilon_i, \varepsilon_j) = 0. \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2\bar{x}}{n} \sum_{i=1}^n Var[\varepsilon_i] \omega_i, \\
&\quad \downarrow \text{ or } Var[\varepsilon_i] = \sigma^2 \text{ et } \sum_{i=1}^n \omega_i = 0 \\
Var[\hat{\beta}_0] &= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

On observe que la variance des estimateurs dépend de la variance inconnue σ^2 de la distribution des données y_i .

Il nous faudra trouver une estimation de cette variance.

3.5 Estimation de la variance σ^2

Dans le cas du modèle linéaire simple, nous pouvons obtenir deux estimations de cette variance en se concentrant sur les résidus du modèle.

Si on procède à l'estimation de σ^2 par maximum de vraisemblance, nous pouvons montrer qu'un estimateur est donné

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2.$$

Mais cette estimateur là est biaisé. Une version débiaisée de cette estimateur est donnée par l'expression.

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2.$$

Il faut donc diviser la somme des carrés des résidus ε_i par $n-2$ et non par n . Si pour le moment le fait de diviser par $n-2$ n'est pas expliquée, nous pourrions noter que le 2 correspond au nombre de paramètres dans notre modèle de régression β_0 et β_1 .

Nous justifierons ce facteur là lorsque nous traiterons le cas général, *i.e.*, lors de la présentation du modèle linéaire multiple.

Déterminer l'espérance et la variance de ces estimateurs va nous permettre de voir si le modèle appris est intéressant, *i.e.* **sommes nous capable de prédire correctement les valeurs de variable Y à l'aide des valeurs de X et de cette relation linéaire.**

3.6 Mesure du lien entre la variable explicative et la variable à expliquer

Nous avons déjà introduit une mesure permettant d'étudier le lien entre deux variables aléatoires X et Y , la **covariance**. Mais la valeur de cette dernière dépend de l'ordre de grandeur des valeurs prises par les différentes variables aléatoires. Ainsi, pour mesurer le lien entre deux variables aléatoires, on calcule le **coefficient de corrélation linéaire**.

Définition 3.1: Coefficient de corrélation linéaire

Soient X et Y deux variables aléatoires admettant des moments d'ordre 2. On appelle coefficient de corrélation linéaire entre les variables X et Y la quantité ρ définie par

$$\rho = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}.$$

Si $|\rho|$ est proche de 1, on dira que la corrélation entre les deux variables est *forte*, à l'inverse, si elle est proche de 0, elle sera *faible*.

De plus, une valeur **négative** de ρ signifie que, globalement, valeurs *croissantes* de X impliquent des valeurs *décroissantes* de Y (et réciproquement), *i.e.* la pente de la droite de régression sera **négative**. De même, une valeur positive de ρ signifie que des valeurs *croissantes* de X impliquent des valeurs *croissantes* de Y (et réciproquement), *i.e.* la pente de la droite de régression sera **positive**.

3.7 Lien entre la pente du modèle $\hat{\beta}_1$ et le coefficient de corrélation $\hat{\rho}$

On rappelle que l'estimateur $\hat{\beta}_1$ de la pente de notre modèle de régression et l'estimateur du coefficient de corrélation $\hat{\rho}$ sont respectivement donnés par

$$\hat{\beta}_1 = \frac{s_{X,Y}^2}{s_X^2} \quad \text{et} \quad \hat{\rho} = \frac{s_{X,Y}^2}{s_X s_Y},$$

où $s_{X,Y}^2$ désigne la covariance empirique entre X et Y et s_X et s_Y désignent respectivement les écart-types des variables aléatoires X et Y .

Le coefficient de corrélation linéaire $\hat{\rho}$ est directement relié à la pente de la droite de régression $\hat{\beta}_1$ par la relation

$$\hat{\beta}_1 = \hat{\rho} \sqrt{\frac{s_Y^2}{s_X^2}},$$

où $\hat{\rho}$ désigne l'estimateur empirique du coefficient de corrélation, s_X^2 et s_Y^2 sont les estimateurs de la variances des variables aléatoires X et Y . Ainsi, nous pourrions indifféremment tester la force du lien entre les deux variables aléatoires X et Y ou tester la significativité de la pente du modèle.

3.8 Significativité du modèle

Étant donnée la remarque précédemment formulée nous pouvons donc nous intéresser indifféremment à une des deux quantités.

Significativité de la pente On effectue le test afin de savoir si la pente est significativement différente de 0. Les hypothèses sont les suivantes :

$$H_0 : \beta_1 = 0 \quad \text{v.s.} \quad h_1 = \beta_1 \neq 0$$

Notre estimateur de la pente, $\hat{\beta}_1$, suit une distribution normale, tout comme la variable aléatoire Y . Les paramètres de cette loi, calculés en Section 3.4 permettent d'écrire que

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Donc

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{N}(0, 1).$$

De plus, d'après ce que nous avons vu en Section 3.5, un estimateur de la variance est donnée par

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Donc

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2.$$

Ainsi

$$t_{\text{test}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \frac{1}{n-2}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \frac{1}{n-2}}} \underset{\text{sous } H_0}{=} \frac{\frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \frac{1}{n-2}}} \sim T_{n-2}.$$

On rappelle que l'on est amené à rejeter H_0 , au seuil de significativité $\alpha \in [0, 1]$ si la statistique de test t_{test} se trouve en dehors de l'intervalle de confiance de niveau $1 - \alpha$, *i.e.* si

$$t_{\text{test}} \notin [t_{\alpha/2, n-2}, t_{1-\alpha/2, n-2}].$$

On pourrait aussi comparer la p -valeur $= 2\mathbb{P}[T \geq |t_{\text{test}}|]$ au seuil de significativité α et rejeter H_0 dans le cas où la p -valeur est inférieure à ce seuil.

Remarque : nous verrons ultérieurement pourquoi le quotient étudié ci-dessus nous donne bien une loi de Student T_{n-2} .

Significativité de la corrélation On cherche maintenant à effectuer la même analyse mais en étudiant cette fois-ci la significativité du coefficient de corrélation ρ .

Le test nous amène à poser les hypothèses suivantes :

$$H_0 : \rho = 0 \quad \text{v.s.} \quad H_1 : \rho \neq 0.$$

Il repose sur une statistique de test similaire à celle de la statistique de test associée à la pente :

$$t_{\text{test}} = \frac{\hat{\rho} - \rho}{\sqrt{\frac{1 - \hat{\rho}^2}{n - 2}}} \underset{\text{sous } H_0}{=} \frac{\hat{\rho}}{\sqrt{\frac{1 - \hat{\rho}^2}{n - 2}}} \sim T_{n-2}.$$

On procédera de la même façon que précédemment pour conclure sur la significativité de la pente étant donné un seuil de significativité α .

Dans la section suivante (Section 4, nous allons tenter d'expliquer l'origine des degrés de liberté qui interviennent dans la construction des tests statistiques et des intervalles de confiance précédents.

3.9 Ecriture du modèle sous forme matricielle

Remarquons enfin que l'on aurait pu écrire notre modèle de régression linéaire simple sous la forme matricielle suivante :

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

où

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix},$$

Nous utiliserons cette écriture dans lors de présentation du modèle linéaire multiple, *i.e.* lorsque le modèle employé utilise plusieurs descripteurs (ou covariables) X_1, X_2, \dots, X_p , objet de la Section 5.

4 Vecteur gaussien et géométrie des modèles linéaires

Cette section est plus abstraite mais elle se propose de fournir une explication quant aux différents tests employés pour évaluer la qualité d'un modèle mais aussi les degrés de libertés des lois de probabilités associés.

Pour cela, nous aurons besoin des outils d'algèbre linéaire et de leurs interprétations géométriques.

4.1 Vecteurs gaussiens

Définir proprement ce qu'est un vecteur gaussien serait faire appel à des notions de probabilités dont on ne dispose pas à ce stade et qui seront inutiles pour la suite de ce cours. On va se contenter de présenter ce qu'est un **vecteur aléatoire** afin de donner une définition simpliste de ce qu'est un vecteur gaussien.

Vecteur aléatoire . Un vecteur aléatoire X est un vecteur dont les différentes composantes X_1, X_2, \dots, X_d sont des variables aléatoires.

Pour chacune de ses composantes, on peut donc calculer une espérance (si cette dernière existe). Le vecteur de ces espérances s'appellera espérance du vecteur aléatoire, *i.e.*

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

La même remarque reste valable pour les matrices de variable aléatoire. Un point très intéressant (et facile à démontrer) est que la propriété de linéarité de l'espérance reste valable dans le cas des vecteurs aléatoires, *i.e.*

$$\forall \mathbf{A} \in \mathbb{R}^{n \times d}, \forall \mathbf{c} \in \mathbb{R}^d, \quad \mathbb{E}[\mathbf{A}X + \mathbf{c}] = \mathbf{A} \mathbb{E}[X] + \mathbf{c}.$$

On peut aussi regarder ce qui se passe pour la covariance. Cette notion est déjà apparue lorsque nous cherchions à étudier les liens entre deux variables aléatoires et elle peut être étendue à plus que deux variables. On peut ainsi définir une **matrice de covariance** de notre vecteur aléatoire X :

$$\text{Var}[X] = \begin{pmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_d] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \cdots & \text{Cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_d, X_1] & \text{Cov}[X_d, X_2] & \cdots & \text{Var}[X_d] \end{pmatrix}$$

Une façon matricielle de calculer cette matrice est d'écrire :

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top]$$

Vecteurs gaussiens Par la suite, nous appellerons **vecteur gaussien**, un vecteur aléatoire X dont les composantes sont des *variables aléatoires gaussiennes indépendantes*.

Ce sont pas les vecteurs gaussiens en tant que tel qui nous intéresse par la suite, même si nous serons amenés à en étudier (indirectement) lorsque l'on cherchera étudier les propriétés des estimateurs des coefficients de la régression linéaire. Regardons tout de suite résultat fondamental concernant les vecteurs gaussien : le Théorème de Cochran !

Théorème 4.1: Théorème de Cochran

Soit X un vecteur gaussien de \mathbb{R}^n de loi $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ où $\sigma^2 > 0$. Supposons que \mathbb{R}^n se décompose comme la somme directe des sous-espaces vectoriels orthogonaux E_k où $\dim(E_k) = d_k > 0$, $\forall k = 1, \dots, p$, i.e.,

$$\mathbb{R}^n = \bigoplus_{k=1, \dots, p}^\perp E_k.$$

On notera également \mathbf{P}_k la matrice de projection orthogonale sur l'espace E_k et Y_k la projection orthogonale de X sur l'espace E_k .

Alors

- (i) les projections Y_k sont des vecteurs gaussiens indépendants et

$$Y_k \sim \mathcal{N}(\mathbf{P}_k \boldsymbol{\mu}, \sigma^2 \mathbf{P}_k).$$

- (ii) les variables aléatoires $\|Y_k - \mathbf{P}_k \boldsymbol{\mu}\|^2$ sont indépendantes. De plus

$$\frac{1}{\sigma^2} \|Y_k - \mathbf{P}_k \boldsymbol{\mu}\|^2 \sim \chi_{d_k}^2.$$

On ne cherchera pas à démontrer ce résultat dans le cadre de ce cours, mais il nous sera utile pour comprendre la notion de degré de liberté que l'on voit apparaître dans les tests statistiques mais aussi pourquoi on divise par certaines valeurs dans l'estimation d'une variance.

Regardons de suite une application à la variance.

4.2 Application à la variance

On peut essayer de comprendre pourquoi on parle de degré de liberté pour le χ^2 en regardant le problème géométriquement. Prenons un vecteur gaussien X ayant n composantes d'espérance $\boldsymbol{\mu} \in \mathbb{R}^n$ et de variance $\sigma^2 \mathbf{I}$. On peut considérer l'espace vectoriel V engendré par le vecteur $\mathbf{v} \in \mathbb{R}^n$ défini par

$$\mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

et la projection du vecteur gaussien X sur cet espace vectoriel est définie par

$$\mathbf{p}_V(X) = \frac{\langle \mathbf{v}, X \rangle}{\|\mathbf{v}\|^2} \mathbf{v} = \frac{1}{n} \sum_{i=1}^n X_i \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \bar{X} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Si on regarde maintenant l'espace orthogonal à V , que l'on note V^\perp , et que l'on s'intéresse à la projection du vecteur X sur ce dernier, il s'agit du vecteur X auquel on enlève sa projection sur V , c'est à dire :

$$\mathbf{p}_{V^\perp}(X) = X - \bar{X} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}$$

De plus comme V et V^\perp sont des espaces vectoriels orthogonaux, et bien les deux vecteurs de projection $\mathbf{p}_V(X)$ et $\mathbf{p}_{V^\perp}(X)$ sont indépendants.

Cela veut dire en particulier que \bar{X} (obtenu à partir de $\mathbf{p}_V(X)$) et $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ (obtenu à partir de $\mathbf{p}_{V^\perp}(X)$) sont indépendants. C'est une conséquence directe du théorème de Cochran.

De plus, la loi du $\chi^2(\nu)$ est très reliée à ces notions géométriques. En effet, la loi du $\chi^2(\nu)$ est par définition la loi de la somme des carrés de ν variables Gaussiennes $\mathcal{N}(0, 1)$ (voir Définition 4.3). De plus, si on prend un espace vectoriel W de dimension d , la projection de X sur W est encore un vecteur gaussien (ce qu'on peut montrer en utilisant une notion un peu avancée comme la fonction génératrice des moments) et sa norme au carré est la somme des carrés des composantes de cette projection. Comme la dimension est d , on a d composantes (et par changement de variable sur W , on peut les rendre indépendantes). La norme au carré de la projection sur un espace vectoriel de

dimension d a donc une loi du χ^2 à d degrés de liberté. C'est pour cela que la variance empirique corrigée, en particulier, définie par

$$S_X^{c^2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

et qui est la projection d'un vecteur gaussien standard sur un espace vectoriel de dimension $n-1$, satisfait

$$\begin{aligned} (n-1)S_X^{c^2} &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \|P_{V^\perp}(X)\|_2^2 \end{aligned}$$

et suit donc une loi du χ^2 à $n-1$ degrés de liberté.

4.3 Définition des lois de probabilités

Cette section résume les définitions des lois de probabilités utilisés dans le cadre de ce cours. Il n'y a pas d'objectifs précis si ce n'est de faire quelques rappels afin de mieux comprendre l'origine des distributions de certaines variables aléatoires qui entrent en jeu.

Définition 4.1: Loi normale

La loi normale est caractérisée par sa moyenne (ou espérance) μ et par sa variance σ^2 . Cela veut dire que la seule connaissance de ces deux paramètres permet de caractériser intégralement cette loi. Elle admet une densité f définie pour tout nombre réel $x \in \mathbb{R}$ par

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[- \left(\frac{x - \mu}{\sigma} \right)^2 \right].$$

Dans le cas multi-dimensionnel, la densité est caractérisée par un vecteur de moyenne $\boldsymbol{\mu} \in \mathbb{R}^n$ et une matrice de covariance $\boldsymbol{\Sigma} \in \mathcal{M}_n(\mathbb{R})$.

Définition 4.2: Loi du χ^2

Soient X_1, X_2, \dots, X_k , k variables aléatoires indépendantes suivant une loi normale centrée réduite. Alors la variable aléatoire X définie par $X = \sum_{i=1}^k X_i^2$ suit une loi du χ^2 à k degrés de liberté, notée \mathcal{X}_k^2 .

Définition 4.3: Loi de Student T_k

Considérons X une variable aléatoire centrée réduite et U une variable aléatoire suivant une loi du χ_k^2 , *i.e.* du Khi-deux à k degrés de libertés, indépendantes. Alors la variable aléatoire $\mathcal{T}_k = \frac{X}{\sqrt{U/k}}$ suit une loi de Student à k degrés de liberté.

Définition 4.4: Loi de Fisher \mathcal{F}_{k_1, k_2}

Soient U_1 et U_2 deux variables aléatoires indépendantes suivant une loi du Khi-deux à respectivement k_1 et k_2 degrés de liberté. Alors la variable aléatoire $X = \frac{U_1/k_1}{U_2/k_2}$ suit une loi de Fisher, notée $F(k_1, k_2)$ (à k_1 et k_2 degrés de liberté)

5 Modèle Linéaire Gaussien Multiple

Dans cette section, on va supposer que le nombre d'exemples n est toujours plus grand que le nombre de descripteurs (ou variables) $p + 1$ ¹.

Le modèle de régression linéaire multiple s'écrit sous la forme

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

où

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1,1} & x_{n-1,2} & \dots & x_{n-1,p} \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix},$$

et $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1}, \beta_p) \in \mathbb{R}^{p+1}$ est notre vecteur des paramètres du modèle. Le vecteur $\mathbf{y} \in \mathbb{R}^n$ est le vecteur dont on cherche à expliquer les valeurs, la matrice $\mathbf{X} \in \mathbb{R}^{n \times p+1}$ est la matrice explicative. On l'appelle aussi *matrice de design* et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ est le vecteur des erreurs associées à chaque exemple qui sont supposées gaussiennes.

Hypothèses . On formule généralement les hypothèses suivantes pour l'étude du modèle linéaire gaussien

1. le modèle est supposé identifiable, *i.e.* il existe un seul et unique vecteur $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ tel que $\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$. Cela est équivalent au fait que les colonnes de la matrice X sont linéairement indépendantes, *i.e.*, que le rang de la matrice \mathbf{X} est égal à $p + 1$.
2. nos données sont *i.i.d.* comme dans le cas de la régression linéaire simple.
3. les erreurs sont supposées centrées donc $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$
4. les erreurs ont toute la même variance et sont indépendantes, ainsi $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$, ou, de façon équivalente $\text{Var}[\mathbf{y}] = \sigma^2 \mathbf{I}_n$

5.1 Estimation par la méthode des moindres carrés

Regardons de suite l'expression de l'estimateur des moindres carrés.

1. Le cas où $p + 1$ est bien plus grand que n nous pousserait à faire de la statistiques en grande dimension, ce qui n'est pas l'objet de ce cours d'autant que cela rendrait l'étude des modèles plus complexe.

Proposition 5.1: Solution de la régression multiple

Considérons le modèle

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

avec la même signification que précédemment. Si le modèle est bien identifiable, *i.e.*, si la matrice \mathbf{X} est de rang égal $p + 1$, alors la matrice $\mathbf{X}^\top \mathbf{X}$ est inversible et l'estimateur des moindres carrés de $\boldsymbol{\beta}$, solution du problème

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

est donné par

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Démonstration. Concernant l'identifiabilité du modèle, on rappelle que si toutes les colonnes de la matrice \mathbf{X} sont indépendantes, alors $\text{Ker}(\mathbf{X}) = \mathbf{0} = \text{Ker}(\mathbf{X}^\top \mathbf{X})$, or $\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{(p+1) \times (p+1)}$, ainsi $\text{rg}(\mathbf{X}^\top \mathbf{X}) = p + 1$, la matrice est de rang plein et est donc inversible.

Pour déterminer l'expression de l'estimateur *vbêta*, nous allons dériver notre problème et chercher les points critiques puis déterminer leur nature en étudiant la hessienne associée.

Les extrema de la fonction la fonction $\boldsymbol{\beta} \mapsto \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ sont données en cherchant l'endroit où le gradient de cette dernière s'annule. On va donc chercher les valeurs de $\boldsymbol{\beta}$ telles que

$$\frac{\partial}{\partial \boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 0 \iff -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \quad (3)$$

En dérivant à nouveau fonction, on trouve

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = 2\mathbf{X}^\top \mathbf{X} \succ 0,$$

i.e. la matrice hessienne est définie positive, ce qui est le cas ici car il s'agit de la matrice de variance-covariance des données, cette convexité nous permettra de dire que la vecteur $\boldsymbol{\beta}$ vérifiant l'équation (3) est bien solution de notre problème de minimisation.

Or

$$\frac{\partial}{\partial \beta} \|y - X\beta\|_2^2 = 0 \iff -2X^\top(Y - X\beta) = 0,$$

↓ on peut diviser par -2

$$\iff X^\top(y - X\beta) = 0,$$

$$\iff X^\top y - X^\top X\beta = 0,$$

$$\iff X^\top X\beta = X^\top y,$$

↓ le modèle est identifiable donc la matrice $X^\top X$ est inversible

$$\frac{\partial}{\partial \beta} \|y - X\beta\|_2^2 = 0 \iff \beta = (X^\top X)^{-1} X^\top y.$$

□

Tout comme dans le cas du modèle linéaire simple, les prédictions \hat{y} sur les données X utilisées pour estimer β sont définies par

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

Etude des prédictions et erreurs du modèles

D'après ce qui précède, nous avons $\hat{y} = X(X^\top X)^{-1} X^\top y$. Dans cette expression, notons que la matrice $H = X(X^\top X)^{-1} X^\top$ est la matrice de projection du vecteur y sur l'espace engendré par les colonnes de la matrice X , qui est donc un espace de dimension $(p + 1)$.

En écrivant ainsi $\hat{y} = Hy$ et en se rappelant que les erreurs du modèles ε sont définies comme étant la différence entre la prédiction \hat{y} et la vraie valeur y , nous avons

$$\begin{aligned} \varepsilon &= \hat{y} - y, \\ &= y - X\hat{\beta}, \\ &= y - Hy, \\ \varepsilon &= (I_n - H)y. \end{aligned}$$

Ainsi, la matrice $L = I_n - H$ définie également une projection, c'est la projection sur l'espace orthogonal à l'espace engendré par les colonnes de X .

Ainsi, les erreurs du modèle appris ne sont que la projection du vecteur y sur le sous-espace orthogonal engendré par les colonnes de X , un espace de dimension $n - (p + 1)$. Cela est une conséquence du fait que les deux projections étudiées définissent des sous-espaces supplémentaires

Ces remarques seront importantes pour la construction des tests associés à la régression multiple. Mais regardons d'abord quelques bonnes propriétés statistiques de notre estimateur.

Propriétés de l'estimateur

Comme dans le cas du modèle linéaire simple, on va chercher à étudier les propriétés de l'estimateur $\hat{\beta}$.

Proposition 5.2: Propriétés de l'estimateur $\hat{\beta}$

L'estimateur $\hat{\beta}$ des moindres carrés ordinaires est

- (i) un estimateur sans biais du paramètre β , i.e. $\mathbb{E}[\hat{\beta}] = \beta$
- (ii) sa variance est égale à $\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.
- (iii) De plus, $\hat{\beta}$ est l'estimateur **sans biais de variance minimale** parmi tous les estimateurs linéaires sans biais de β .

Démonstration. On rappelle que l'on a

$$\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta \quad \text{et} \quad \text{Var}[\mathbf{y}] = \sigma^2 \mathbf{I}_n.$$

L'estimateur des moindres carrés ordinaires est donné par $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{A}\mathbf{y}$, où $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.

(i) Espérance de l'estimateur : $\mathbb{E}[\hat{\beta}]$ On repart de la définition de l'estimateur dans la Proposition 5.1 et on garde l'esprit que \mathbf{X} est **déterministe** (ce n'est pas une variable aléatoire), la seule partie aléatoire dans l'expression de $\hat{\beta}$ vient de la variable aléatoire Y_i via la variable aléatoire ε_i .

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}], \\ &\quad \downarrow \text{seule } \mathbf{y} \text{ est aléatoire} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{y}], \\ &\quad \downarrow \text{par définition de } \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{X}\beta + \varepsilon], \\ &\quad \downarrow \text{seule } \varepsilon \text{ est aléatoire} \end{aligned}$$

$$\begin{aligned}
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\varepsilon}], \\
&\quad \downarrow \text{ car } \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \text{ hypothèse sur les erreurs du modèle} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}, \\
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}.
\end{aligned}$$

(ii) **Variance de l'estimateur** $\text{Var}[\hat{\boldsymbol{\beta}}]$ On garde à l'esprit ce que nous avons utilisés précédemment, à savoir que seule $\boldsymbol{\varepsilon}$, *i.e.* les erreurs sont aléatoires ainsi que les propriétés de la variance.

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top], \\
&\quad \downarrow \text{ On repart de la définition de } \hat{\boldsymbol{\beta}} \\
&= \mathbb{E}\left[((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta})(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta})^\top\right], \\
&\quad \downarrow \text{ définition de } \mathbf{y} \\
&= \mathbb{E}\left[((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta})(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta})^\top\right], \\
&\quad \downarrow \text{ on développe et on simplifie les expressions} \\
&= \mathbb{E}\left[((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon})(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon})^\top\right], \\
&\quad \downarrow \text{ par définition de la transposition} \\
&= \mathbb{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}\right], \\
&\quad \downarrow \text{ seule la partie en } \boldsymbol{\varepsilon} \text{ est aléatoire} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}, \\
&\quad \downarrow \text{ or } \mathbb{E}[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top] = \sigma^2 \mathbf{I}_n \\
&= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}, \\
&\quad \downarrow \text{ par simplification} \\
\text{Var}[\hat{\boldsymbol{\beta}}] &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.
\end{aligned}$$

(iii) **Estimateur linéaire de plus petite variance** Considérons un autre estimateur linéaire $\tilde{\boldsymbol{\beta}}$ sans biais de $\boldsymbol{\beta}$ tel qu'il existe une matrice \mathbf{B} telle que $\tilde{\boldsymbol{\beta}} = \mathbf{B} \mathbf{y} = \boldsymbol{\beta}$. On a alors $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \mathbf{B} \mathbb{E}[\mathbf{y}]$. Or l'espérance de \mathbf{y} est égale à $\mathbf{X} \boldsymbol{\beta}$, on a donc $\mathbf{B} \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$ donc

$$\mathbf{B} \mathbf{X} = \mathbf{I}_{p+1}.$$

Cela nous donne

$$\mathbf{B}^\top \mathbf{A} = \mathbf{B} \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} = \mathbf{A}^\top \mathbf{A}.$$

En posant $\mathbf{C} = \mathbf{B} - \mathbf{A}$, on trouve $\mathbf{C} \mathbf{A}^\top = \mathbf{0}$ et $\mathbf{A}^\top \mathbf{C} = \mathbf{0}$.
Ainsi la variance de $\tilde{\boldsymbol{\beta}}$ est donnée par

$$\begin{aligned} \text{Var}[\tilde{\boldsymbol{\beta}}] &= \sigma^2 \mathbf{B} \mathbf{B}^\top, \\ &\quad \downarrow \text{par définition de } \mathbf{C} \\ &= \sigma^2 (\mathbf{A} + \mathbf{C}) (\mathbf{A} + \mathbf{C})^\top, \\ &\quad \downarrow \mathbf{C} \mathbf{A}^\top = \mathbf{0} \text{ et } \mathbf{A}^\top \mathbf{C} = \mathbf{0} \\ &= \sigma^2 \mathbf{A} \mathbf{A}^\top + \sigma^2 \mathbf{C} \mathbf{C}^\top, \\ &= \text{Var}[\hat{\boldsymbol{\beta}}] + \sigma^2 \mathbf{C} \mathbf{C}^\top. \end{aligned}$$

Cette dernière matrice est $\mathbf{C} \mathbf{C}^\top$ est semi-définie positive, ces valeurs propres sont donc toutes positives, donc pour tout vecteur $\mathbf{u} \in \mathbb{R}^{p+1}$, nous avons

$$\mathbf{u}^\top \text{Var}[\tilde{\boldsymbol{\beta}}] \mathbf{u} = \mathbf{u}^\top \text{Var}[\hat{\boldsymbol{\beta}}] \mathbf{u} + \mathbf{u}^\top \mathbf{C} \mathbf{C}^\top \mathbf{u} = \mathbf{u}^\top \text{Var}[\hat{\boldsymbol{\beta}}] \mathbf{u} + \underbrace{\|\mathbf{C}^\top \mathbf{u}\|_2^2}_{\geq 0} \geq \mathbf{u}^\top \text{Var}[\hat{\boldsymbol{\beta}}] \mathbf{u}.$$

On en déduit que $\text{Var}[\tilde{\boldsymbol{\beta}}] \geq \text{Var}[\hat{\boldsymbol{\beta}}]$ pour la relation d'ordre précédemment définie. \square

5.2 Estimation par maximum de vraisemblance

On peut également trouver l'expression des estimateurs par maximum de vraisemblance en se rappelant que pour tout $i \in \llbracket 1, n \rrbracket$

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Ainsi la variable aléatoire y_i est distribuée selon une loi normale centrée en $\mathbf{x}_i^\top \boldsymbol{\beta}$ et de variance σ^2 .

Nos données étant supposées indépendantes, la loi jointe de \mathbf{y} est égale au produit des lois y_i .

La densité jointe évaluée en les observations est donnée par

$$f(\mathbf{y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right).$$

Dans cette densité jointe, deux quantités sont inconnues : (i) le vecteur $\boldsymbol{\beta}$ des paramètres et (ii) la variance des erreurs σ^2 . On va donc étudier la fonction de vraisemblance \mathcal{L} qui est une fonction des paramètres inconnues

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right).$$

Simplifions cette expression

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right), \\ &\quad \downarrow \text{ en utilisant la propriété de l'exponentielle} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right). \end{aligned}$$

Or on se rappelle que si $\mathbf{u} \in \mathbb{R}^n$, alors $\sum_{i=1}^n u_i^2 = \|\mathbf{u}\|_2^2$, que l'on va appliquer à ce qui figure au sein de l'exponentielle. Ainsi

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right), \\ &\quad \downarrow \text{ or } \sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} = \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right). \end{aligned}$$

où les notations \mathbf{X} et \mathbf{y} sont celles employées dans la partie précédente.

On se rappelle que les meilleurs paramètres sont ceux qui permettent de maximiser la vraisemblance de nos données. Mais cette expression est bien trop complexe à manipuler. Donc au lieu de maximiser la vraisemblance L on va chercher à maximiser la log-vraisemblance ℓ , définie par $L = \ln(L)$ soit

$$\ell(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{2\sigma^2}.$$

Les valeurs de σ^2 et β qui maximisent la vraisemblance sont données en résolvant les équations d'Euler définies par

$$\frac{\partial \ell}{\partial \beta}(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) = 0 \quad \text{et} \quad \frac{\partial \ell}{\partial \sigma^2}(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) = 0.$$

Concentrons nous sur la première équation, ce qui nous donne

$$\begin{aligned} & \frac{\partial \ell}{\partial \beta}(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) = 0, \\ & \downarrow \text{on dérive notre norme} \\ & \iff -\frac{\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} = 0, \\ & \iff -\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\beta = 0, \\ & \downarrow \text{on isole le vecteur } \beta \\ & \iff \beta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \end{aligned}$$

Faisons de même avec la deuxième équation en utilisant l'estimateur $\hat{\beta}$ précédemment obtenu et en notant que

$$\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

On a ainsi,

$$\begin{aligned} & \frac{\partial \ell}{\partial \sigma^2}(\beta, \sigma^2; \mathbf{y}, \mathbf{X}) = 0, \\ & \downarrow \text{on en utilisant les notations précédentes} \\ & \iff -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2}{(\sigma^2)^2} = 0, \\ & \downarrow \text{en multipliant par } 2(\sigma^2)^2 \\ & \iff -n\sigma^2 + \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 = 0, \\ & \downarrow \text{on isole } \sigma^2 \\ & \iff \sigma^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2}{n}. \end{aligned}$$

Ainsi les estimateurs obtenus par maximum de vraisemblance sont donnés par

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \hat{\beta})^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2}{n}.$$

Remarque L'estimateur de σ^2 ainsi obtenu est biaisé ! Il faudrait corriger cet estimateur pour le rendre non biaisé.

5.3 Estimation de la variance σ^2

Le paramètre σ^2 qui n'est rien d'autre que la variance de nos résidus est défini par

$$\sigma^2 = \text{Var}[\epsilon] = \text{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2].$$

On rappelle que dans un modèle linéaire l'espérance de la variable aléatoire Y est estimé par $\mathbf{X}\hat{\beta}$. On va alors estimer la valeur de σ^2 à l'aide des résidus $\hat{\epsilon}_i$ de notre modèle. L'estimateur $\hat{\sigma}^2$ de la variance de nos résidus est défini par

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

Proposition 5.3: Propriété de l'estimateur de la variance

Soit $\hat{\sigma}^2$, l'estimateur de la variance σ^2 des résidus défini par

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

Cet estimateur est un estimateur **sans biais** de σ^2 .

Cette proposition vient compléter la Proposition 5.2.

Démonstration. En notant $\hat{y}_i = \mathbf{x}_i^\top \hat{\beta}$, on peut écrire que $\sum_{i=1}^n \hat{\epsilon}_i^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{L}\mathbf{y}\|^2$, où \mathbf{L} est la matrice de projection des observations \mathbf{y} sur l'espace orthogonale à l'espace engendré par les données \mathbf{X} .

On a donc

$$\begin{aligned}
\mathbb{E}[\|\mathbf{y} - \hat{\mathbf{y}}\|^2] &= \mathbb{E}[\|\mathbf{L}\mathbf{y}\|^2], \\
&\downarrow \text{définition de la norme} \\
&= \mathbb{E}[\text{Tr}((\mathbf{L}\mathbf{y})^\top (\mathbf{L}\mathbf{y}))], \\
&= \mathbb{E}[\text{Tr}(\mathbf{y}^\top \mathbf{L}^\top \mathbf{L} \mathbf{y})], \\
&\downarrow \text{par définition de } \mathbf{L} \\
&= \mathbb{E}[\text{Tr}(\mathbf{y}^\top \mathbf{L} \mathbf{y})], \\
&\downarrow \text{Linéarité de la trace + propriété de la trac} \\
&= \text{Tr}(\mathbb{E}[\mathbf{L} \mathbf{y} \mathbf{y}^\top]), \\
&\downarrow \text{linéarité de l'espérance} \\
&= \text{Tr}(\mathbf{L} \mathbb{E}[\mathbf{y} \mathbf{y}^\top]), \\
&\downarrow \text{définition de la variance} \\
&= \text{Tr}(\mathbf{L} \text{Var}[\mathbf{y}]).
\end{aligned}$$

Or $\text{Var}[\mathbf{y}] = \sigma^2$, c'est une hypothèse formulée dans le cadre du modèle linéaire gaussien. Ainsi

$$\mathbb{E}\left[\sum_{i=1}^n \hat{\varepsilon}_i^2\right] = \mathbb{E}[\|\mathbf{y} - \hat{\mathbf{y}}\|^2] = \sigma^2 \text{Tr}(\mathbf{L}) = \sigma^2(n - p - 1),$$

car \mathbf{L} est un projecteur sur un espace de dimension $n - p - 1$ (dimension du sous espace orthogonal à l'espace engendré par les données $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$). \square

5.4 Test de nullité d'un coefficient de la régression

On souhaite déterminer si le j -ème coefficient de la régression est significativement différent de 0 ou non. En d'autres termes, on va chercher à déterminer si la j -ème variable permet bien d'expliquer, en partie, les valeurs prises par la variable aléatoire Y . On pose donc les hypothèses suivantes :

$$H_0 : \beta_j = 0 \quad \text{v.s.} \quad H_1 : \beta_j \neq 0.$$

Pour construire ce test, nous avons besoin de connaître la loi $\hat{\beta}_j$ pour en déduire la loi de notre statistique de test sous H_0 .

Pour cela, on peut montrer le résultat suivant :

Théorème 5.1: Résultats sur les estimateurs

Soient $\hat{\beta}$ et $\hat{\sigma}^2$ les estimateurs de β et σ^2 où $\hat{\beta}$ est obtenu par maximum de vraisemblance.

Alors $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.

- l'estimateur $\hat{\beta}$ suit une loi gaussienne multivariée de moyenne β et de matrice de variance $(\mathbf{X}^\top \mathbf{X})^{-1}$, i.e.,

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}).$$

- la variable aléatoire $\hat{\sigma}^2 \frac{n-p-1}{\sigma^2}$ suit une loi du χ^2 à $n-p-1$ degrés de libertés, i.e.,

$$\hat{\sigma}^2 \frac{n-p-1}{\sigma^2} \sim \chi_{n-p-1}^2.$$

Démonstration. On rappelle que la variable aléatoire \mathbf{y} suit une loi gaussienne multivariée de moyenne $\mathbf{X}\beta$ et de matrice de variance $\sigma^2 \mathbf{I}_n$.

Considérons les projections de \mathbf{y} sur l'espace engendré par les colonnes de la matrice \mathbf{X} , définie par une matrice $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, et la projection sur l'espace orthogonale à l'espace engendré par les colonnes de la matrice \mathbf{X} , définie par la matrice $\mathbf{L} = \mathbf{I}_n - \mathbf{H}$. On a donc les relations

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{H}\mathbf{y}, \\ \hat{\varepsilon} &= \mathbf{L}\mathbf{y}.\end{aligned}$$

S'agissant de deux projections d'un vecteur gaussien sur des sous-espaces supplémentaires orthogonaux, le théorème de Cochran nous assure la résultante des deux projections est encore un vecteur gaussien, qu'ils sont indépendants et que l'on a

$$\begin{aligned}\hat{\mathbf{y}} &\sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{H}), \\ \hat{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{L}).\end{aligned}$$

De plus $\hat{\beta}$ est aussi gaussien comme transformation linéaire d'un vecteur gaussien et nous nous avons $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ où $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$. Il ne nous reste qu'à déterminer l'espérance et la variance de cette dernière.

Nous avons vu $\hat{\beta}$ est un estimateur sans biais de β et que $\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$, ainsi

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}).$$

Enfin, d'après le Théorème de Cochran (Théorème 4.1), on en déduit que

$$\frac{\|\mathbf{L}\mathbf{y}\|^2}{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \sim \chi_{n-p-1}^2.$$

En effet, la matrice H est une matrice de projection sur un sous-espace de dimension $p+1$ (nombres de colonnes de la matrice \mathbf{X}) et $\mathbf{L} = \mathbf{I}_n - \mathbf{H}$ est la matrice projection sur l'orthogonal de l'espace précédent qui est de dimension $n - (p+1)$. \square

Le résultat nous renseigne directement sur la distribution de l'estimateur $\hat{\beta}$. On peut ensuite se servir de cette distribution pour construire un test statistique permettant de tester la significativité d'un paramètre $\beta_j, j \in \llbracket 1, n \rrbracket$.

Corollaire 5.1: Statistiques de test

Pour tout $j \in \llbracket 0, p \rrbracket$, on considère β_j le coefficient de la régression associée à la variable X_j et $\hat{\beta}_j$ son estimateur.

Alors

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim T_{n-(p+1)},$$

où $\sigma_{\hat{\beta}_j}^2$ désigne la racine carrée de la valeur de se trouvant en position $(j+1, j+1)$ de la matrice $\hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

Démonstration. Il nous suffit de d'utiliser la Proposition 5.3 qui nous indique que $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 C_{jj})$ où C_{jj} représente le $j+1$ -ème élément sur la diagonale de la matrice $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Ainsi, la variable aléatoire $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 C_{jj}}}$ suit une loi normale centrée et réduite.

Malheureusement, la valeur de σ^2 est inconnue, mais nous savons qu'un bon estimateur de σ^2 est donné par

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

où les $\varepsilon_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Ainsi nous avons également

$$\hat{\sigma}^2 \frac{n-p-1}{\sigma^2} \sim \chi_{n-p-1}^2.$$

Or les deux variables aléatoires sont indépendantes, donc

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{C_{jj}}}}{\sqrt{\frac{\hat{\sigma}^2 \frac{n-p-1}{\sigma^2}}{n-p-1}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{C_{jj}}} \sim T_{n-p-1},$$

d'après la définition d'une loi de Student. \square

Dans le cadre d'un test bilatéral, où l'on cherche à regarder le coefficient β_j est significativement différent de 0, on sera amené à rejeter l'hypothèse nulle, au risque d'erreur $\alpha \in (0, 1)$, si la statistique de test

$$|t_{\text{test}}| = \left| \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{C_{jj}}} \right| \geq t_{1-\alpha/2, n-p-1}.$$

Ce dernier résultat permet également de construire des intervalles de confiance sur les estimateurs des paramètres du modèle.

5.5 Prédiction et intervalles de prédiction

Une fois le modèle construit et vérifié, *i.e.*, après avoir vérifié la significativité du modèle et des différents coefficients, nous pouvons nous intéresser aux prédictions effectuées par le modèle et s'intéresser notamment aux possibles valeurs prédites par le modèle.

Pour cela, nous allons considérer une nouvelle donnée \mathbf{x}_{new} et on notera la prédiction effectuée par le modèle \hat{y}_{new} .

On peut s'intéresser à deux niveaux de prédiction : *(i)* s'intéresser à la prédiction dite *individuelle* \hat{y}_{new} ou *(ii)* s'intéresser à la valeur moyenne prédite par le modèle pour cette nouvelle donnée, *i.e.* $\mathbb{E}[\hat{y}_{\text{new}}]$.

5.5.1 Intervalle de confiance sur la prédiction

On cherche à construire un intervalle de confiance sur la prédiction pour un individu (et non plus sur l'espérance de la prédiction). On rappelle que pour un individu \mathbf{x}_{new} , la prédiction \hat{y}_{new} effectuée par le modèle est donnée par :

$$\hat{y}_{\text{new}} = \hat{\beta} \mathbf{x}_{\text{new}},$$

On en déduit que l'espérance de la valeur prédite est donnée par

$$\mathbb{E}[\hat{y}_{\text{new}}] = \mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}.$$

La variance associée à cette prédiction est donnée par

$$\begin{aligned} \text{Var}[\hat{y}_{\text{new}}] &= \mathbb{E}[(\hat{y}_{\text{new}} - \mathbb{E}[\hat{y}_{\text{new}}])^2], \\ &= \mathbb{E}[(\mathbf{x}_{\text{new}}^\top (\hat{\beta} - \boldsymbol{\beta}))^2], \\ &= \mathbb{E}[\mathbf{x}_{\text{new}}^\top (\hat{\beta} - \boldsymbol{\beta})(\hat{\beta} - \boldsymbol{\beta})^\top \mathbf{x}_{\text{new}}], \\ &= \mathbf{x}_{\text{new}}^\top \mathbb{E}[(\hat{\beta} - \boldsymbol{\beta})(\hat{\beta} - \boldsymbol{\beta})^\top] \mathbf{x}_{\text{new}}, \\ &= \mathbf{x}_{\text{new}}^\top \text{Var}[\hat{\beta}] \mathbf{x}_{\text{new}}, \\ \text{Var}[\hat{y}_{\text{new}}] &= \sigma^2 \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}}. \end{aligned}$$

On en déduit alors la distribution des valeurs de \hat{y}_{new} pour obtenir le fait que (en utilisant l'indépendance) :

$$\hat{y}_{\text{new}} - y_{\text{new}} \sim \mathcal{N}(0, \sigma^2(1 + \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}})).$$

Or, nous avons vu que la variable aléatoire $\frac{y_{\text{new}} - \hat{y}_{\text{new}}}{\sqrt{\text{Var}[y_{\text{new}} - \hat{y}_{\text{new}}]}}$ suit une loi de Student à $n - p - 1$ degrés de libertés. Donc l'intervalle de confiance de niveau $1 - \alpha$ sur la valeur prédite \hat{y}_{new} est directement donné par

$$\hat{y}_{\text{new}} \pm t_{1-\alpha/2, n-p-1} \hat{\sigma} \sqrt{1 + \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}}}.$$

5.5.2 Intervalle de confiance sur la valeur moyenne prédite

On peut montrer que l'intervalle de confiance de niveau $1 - \alpha$ sur l'espérance de la prédiction, $\mathbb{E}[y_{\text{new}}]$, est donné par :

$$\hat{y}_{\text{new}} \pm t_{1-\alpha/2, n-p-1} \hat{\sigma} \sqrt{\mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}}}.$$

En effet, en reprenant la même démonstration que précédemment, on a

$$\hat{y}_{\text{new}} \sim \mathcal{N}(\mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}, \sigma^2(1 + \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}})).$$

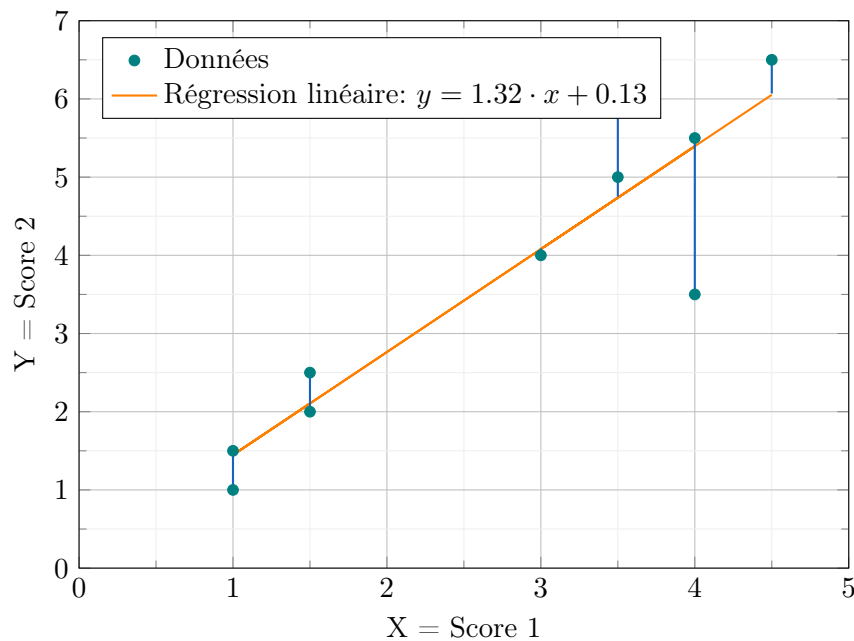
Ainsi la variable aléatoire $\frac{y_{\text{new}} - \mathbf{x}_{\text{new}}^\top \boldsymbol{\beta}}{\sqrt{\text{Var}[y_{\text{new}} - \hat{y}_{\text{new}}]}}$ suit une loi de Student à $n - p - 1$ degrés de libertés, d'où l'intervalle de confiance précédemment cité.

Cet intervalle de confiance est le même que celui obtenu à la section précédente, on aura simplement enlevé la valeur 1 de $1 + \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}}$. Il y a donc une incertitude plus faible sur une prédiction en espérance que sur une prédiction individuelle.

5.6 Qualité du modèle

On rappelle que l'on cherche à estimer les paramètres du modèle de sorte à minimiser la carré de l'écart entre les valeurs observées y_i et les valeurs prédites par les prédites par le modèle \hat{y}_i , *i.e.* à résoudre le problème

$$\min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n \varepsilon_i^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$



La somme des carrés des écarts entre y_i et \hat{y}_i s'appelle aussi la **Somme des carrés résiduels (SCR)**, *i.e.* ;

$$\text{SCR} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \varepsilon_i^2.$$

Elle est très liée à la variance initialement présente dans nos données. En fait, nous pouvons montrer que l'on a la relation suivante entre la variance de nos observations et **SCR** :

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SCR},$$

où \bar{y} désigne la valeur moyenne de $\mathbf{y} \in \mathbb{R}^n$.

On peut voir le terme **SCT** comme la variation (ou quantité d'information) présente dans les données, **la somme des carrés expliqués (SCE)** représente la variation expliquée par le modèle et **SCR** celle qui n'est pas expliquée par le modèle (ou variance résiduelle).

A l'aide de ces différentes quantités, on peut à nouveau construire un test statistique permettant de tester la significativité globale du modèle, *i.e.*, on effectue le test suivant :

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p \quad \text{v.s.} \quad \exists j \in \llbracket 1, p \rrbracket \beta_j \neq 0.$$

Dit autrement, tester la significativité du modèle revient à tester l'hypothèse selon laquelle aucune des covariables n'explique les valeurs observées y . Regardons comment est construit ce test.

Analyse de Variance et significativité du modèle Les termes précédents représentant des variances, on est souvent amené à résumer les informations de notre modèle dans une table d'analyse de variance

Table d'analyse de variance			
Source de Variation	Somme des carrés	Degrés de libertés	Carrés moyens
Modèle (SCE)	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	p	$MSE = \frac{SCE}{p}$
Résiduelle (SCR)	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - p - 1$	$MSR = \frac{SCR}{n - p - 1}$
Totale (SST)	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	$MST = \frac{SST}{n - 1}$

Remarquons que les différentes sommes de carrés peuvent aussi s'écrire à l'aide de la norme de deux vecteurs, comme cela a pu être fait pour la partie résiduelle.

On peut alors définir une statistique de test F_{test} suivante pour tester la significativité globale du modèle :

$$F_{\text{test}} = \frac{\text{MSA}}{\text{MSW}} F_{p, n-p-1}.$$

Cette statistique de test suit donc une loi de Fisher à p et $n - p - 1$ degrés de liberté. On rejette alors l'hypothèse H_0 au risque d'erreur $\alpha \in (0, 1)$ si

$$F_{\text{test}} > f_{p, n-p-1, 1-\alpha},$$

i.e. si la statistique de test prend une valeur supérieure au quantile d'ordre $1 - \alpha$ d'une loi de Fisher à p et $n - p - 1$ degrés de libertés.

On remarque que l'on effectue ici un test de rapport de variance *unilatéral supérieur*.

On peut également retrouver ce test en se basant sur les résultats précédemment établi mais notons que cette première vision est plus simple que celle présentée ci-dessus.

Proposition 5.4: Test de Fisher - Significativité du modèle

On considère le test statistique suivant :

$$H_0 : \beta = \mathbf{0} \quad \text{v.s.} \quad \exists j \in \llbracket 0, p \rrbracket, \beta_j \neq 0.$$

Alors

$$\frac{(\hat{\beta} - \beta)(\mathbf{X}^\top \mathbf{X})^{-1}(\hat{\beta} - \beta)}{(p+1)\hat{\sigma}^2} \sim F_{p+1, n-p-1}.$$

où $\hat{\sigma}^2$ est un estimateur de la variance des résidus.

On peut en fait montrer un résultat plus général. Considérons une matrice $\mathbf{C} \in \mathbb{R}^{q \times (p+1)}$ de rang $1 \leq q \leq p+1$, alors on a

$$\frac{(\mathbf{C}\hat{\beta} - \mathbf{C}\beta)\mathbf{C}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{C}^\top(\mathbf{C}\hat{\beta} - \mathbf{C}\beta)}{q\hat{\sigma}^2} \sim F_{q, n-p-1}.$$

Démonstration. On se rappelle que l'on a $\hat{\beta} - \beta \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$, ainsi, en multipliant ce vecteur par la matrice \mathbf{C} , on a $\mathbf{C}(\hat{\beta} - \beta) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{C}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{C}^\top)$.

Or la matrice \mathbf{C} est de rang q , donc la matrice $\mathbf{C}^\top \mathbf{C}$ est inversible donc la matrice $\mathbf{C}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{C}^\top$ l'est également.

On en déduit que

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})}{\sigma^2} \sim \chi_q^2.$$

Or, d'après le Théorème de Cochran, on sait que $\hat{\boldsymbol{\beta}} \perp \hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^n \hat{\epsilon}_i^2$,
et

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2.$$

Ainsi

$$\frac{\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})}{\sigma^2}}{\frac{q}{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}}} = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})}{q\hat{\sigma}^2} \sim F_{q,n-p-1},$$

en utilisant la définition d'une loi de Fisher. □

La Proposition 5.4 consiste à prendre \mathbf{C} comme étant la matrice identité d'ordre $p+1$.

La Proposition 5.4 permet également de retrouver le test présenté au paragraphe précédent, il suffit pour cela de supprimer la constante β_0 du modèle de notre vecteur $\boldsymbol{\beta}$.

Ce résultat général est également intéressant car il peut permettre de tester si des "sous-modèles", appelés aussi *modèles réduits*, sont significatifs ou non.

Regardons maintenant comment quantifier la qualité de notre modèle.

Qualité du modèle et ajustement Dans le cas du modèle linéaire simple, on a pu évaluer la qualité de l'ajustement du modèle au donnée en évaluant la corrélation entre les deux variables X et Y . Nous aurons plus de mal à faire cela en dimension supérieure mais il est possible d'évaluer cet ajustement via le **coefficient de détermination** R^2 , qui étudie la part de variance expliquée par le modèle (SCE) par rapport à la variance totale (SCT). Plus précisément

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2}.$$

Ce coefficient de détermination est donc une valeur comprise entre 0 et 1. Plus cette valeur est proche de 1 plus le modèle appris permet d'expliquer les données observées.

Remarque : on pourrait penser qu'avoir une valeur proche de 1 est très intéressante en pratique mais cela n'est pas toujours une garantie que le modèle appris soit fiable et performant. Il se peut que le modèle apprenne par coeur les données et qu'il fasse ce que l'on appelle du *sur-apprentissage*. Nous verrons cela plus tard dans des cours de *Machine Learning*.

Ce critère présente malheureusement un inconvénient, sa valeur augmente naturellement avec le nombre p de variables explicatives présentent dans le modèle. Il est donc parfois d'usage de considérer un autre critère, que l'on appelle le **coefficient de détermination ajusté**, R^2 -ajusté, qui prend en compte le nombre de variables dans le modèle. Cette quantité est définie par

$$R^2\text{-ajusté} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

En plus de permettre l'évaluation de la qualité des modèles, ces critères vont nous permettre de faire de la sélection de modèles ?

5.7 Construction et sélection de modèles

Plusieurs questions se posent lorsque l'on cherche à construire un modèle de prédiction à l'aide des informations dont on dispose :

1. Est-ce que le modèle appris est qualitatif ? C'est un point que l'on a précédemment étudié.
2. Est-ce que les variables utilisées pour construire le modèle apportent de l'information différente ? C'est la question de la redondance de l'information.
3. Est-ce que toutes ces variables permettent de contribuer, de façon significative, aux performances du modèle ? Ici, c'est plutôt l'importance de l'information.

Redondance de l'information La réponse à la deuxième question est importante ne serait-ce que pour assurer la validité du modèle. En effet, on rappelle que pour pouvoir estimer les paramètres β , il faut que la matrice \mathbf{X} soit de rang plein afin que la matrice $\mathbf{X}^\top \mathbf{X}$ soit inversible. S'il existe une variable X_j qui peut s'exprimer comme une combinaison linéaire de toutes les autres variables X_k , alors notre matrice ne sera plus inversible, *i.e.* s'il existe des coefficients α_k tels que

$$X_j = \sum_{k=1, k \neq j}^p \alpha_k X_k.$$

Cette relation signifie aussi que l'on est capable de prédire la valeur de X_j étant donnée les valeurs X_k à l'aide ... d'un modèle linéaire! Partant de cette observation, on pourrait donc aussi estimer que des variables X_j ne sont que *faiblement utiles* (voir qu'elles peuvent nuire à la qualité du modèle) s'il existe un lien linéaire fort avec les autres variables X_k .

Pour évaluer cela en pratique, nous allons ... construire un modèle linéaire entre X_j et les autres variables X_k et évaluer la *qualité de l'ajustement*, le R^2 . Si ce dernier est trop élevé, on estimera que la variable X_j est redondante.

Ce critère est appelé le critère du *VIF* pour *Variation Inflation Factor* et est défini par

$$VIF(X_j) = \frac{1}{1 - R_j^2},$$

où R_j^2 est le coefficient de détermination associé au modèle $X_j = \sum_{k=1, k \neq j}^p \alpha_k X_k$. La variable est alors exclue des données si son *VIF* est supérieur à 10 (certains auteurs font le choix de prendre la valeur 5).

Ainsi, avant de chercher à construire un modèle, on va d'abord chercher à supprimer l'information redondance en appliquant la procédure suivante :

1. Calculer les *VIF* pour l'ensemble des variables X_j en évaluant les R_j^2 associés aux modèles

$$X_j = \sum_{k=1, k \neq j}^p \alpha_k X_k.$$

2. S'il n'y a qu'une seule variable qui présente un *VIF* supérieur à 10, exclure cette variable du jeu de données et terminer la procédure.
3. Si plusieurs variables présentent un *VIF* supérieur à 10, on supprime uniquement la variable ayant le *VIF* le plus élevé et on revient à l'étape 1, jusqu'à ce que toutes les variables aient un *VIF* inférieur à 10.

Il reste maintenant à regarder quelles sont les informations essentielles pour établir notre meilleur modèle.

Sélection de modèle Nous allons maintenant regarder comment sélectionner des modèles de façon générale, *i.e.* quel(s) critère(s) nous pouvons utiliser pour comparer des modèles. Nous regarderons ensuite le cas particulier de modèles emboîtés.

Nous avons déjà vu le critère du R^2 -ajusté précédemment, mais nous pouvons également utiliser des critères basés sur la vraisemblance de nos données.

(i) C_p Mallows coefficient [Gilmour, 1996].

Pour un modèle Ω_q comprenant $q < p$ variables, ce critère est défini par

$$C_p(\Omega_q) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\Omega_q)\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2} 2(q+1) - n,$$

où $\hat{\mathbf{y}}$ désigne les prédictions effectuées avec le modèle complet et $\hat{\mathbf{y}}(\Omega_q)$ les prédictions effectuées avec le modèle réduit Ω_q .

(ii) **Aikaike Information Criterion (AIC)** [**Akaike, 1974**].

Ce premier critère a été essentiellement motivé par l'étude des modèles gaussiens et est défini pour un modèle Ω_q , comprenant $q < p$ variables

$$\text{AIC}(\Omega_q) = n(\ln(2\pi) + 1) + n \ln \left(\frac{\|\mathbf{y} - \hat{\mathbf{y}}(\Omega_q)\|^2}{n} \right) + 2(q+2),$$

où $\hat{\mathbf{y}}(\Omega_q)$ désigne les prédictions effectuées par le modèle à q variables.

Dans le cas gaussien, on peut montrer que les critères AIC et C_p sont équivalents. Le critère suivant est le plus employé en statistique.

(iii) **Bayesian Information Criterion (BIC)** [**Schwarz, 1978**].

Avec les mêmes notations que précédemment, on a

$$\text{BIC}(\Omega_q) = n(\ln(2\pi) + 1) + n \ln \left(\frac{\|\mathbf{y} - \hat{\mathbf{y}}(\Omega_q)\|^2}{n} \right) + \ln(n)(q+2).$$

On peut montrer que lorsque $n > 7$ on a $\ln(n) > 2$, donc le critère BIC aura tendance à sélectionner des modèles plus petits que le critère AIC. L'objectif est de retenir le modèle Ω_q qui **minimise** l'un de ces trois critères.

Il existe une procédure qui permet de tester si une variable permet d'augmenter significativement ou non les performances d'un modèle. On dit que l'on cherche à *comparer des modèles emboîtés*.

Pour cela, considérons un entier $q < p$ et considérons les modèles Ω_q et Ω_{q+1} où Ω_q est un modèle comprenant q variables parmi les $q+1$ du modèle Ω_q . Afin de tester si l'ajout ou le retrait de cette variable a un impact significatif sur les performances du modèles, on peut effectuer le test statistique dont les hypothèses sont les suivantes :

H_0 : le modèle Ω_q est valide v.s. le modèle Ω_{q+1} est valide.

La statistique de test F_{test} que l'on considère est définie par

$$F_{\text{test}} = \frac{\frac{SCR(\Omega_q) - SCR(\Omega_{q+1})}{1}}{\frac{SCR(\Omega_{q+1})}{n - (q+2)}},$$

où $q + 2$ représente le nombre de paramètre du modèle Ω_{q+1} et 1 correspond à la différence, en terme de nombre de paramètres, des modèles Ω_{q+1} et Ω_q .

Cette statistique de test va suivre, sous l'hypothèse H_0 , une loi de Fisher à respectivement 1 et $n - (q + 2)$ degrés de liberté.

5.8 Analyse des résidus et détection d'outliers

On cherche maintenant à vérifier si les hypothèses du modèle linéaire gaussien sont valides ou non. C'est une étape que l'on effectue *a posteriori* après avoir retenu le meilleur modèle selon les critères statistiques définies dans la précédente section.

Nous avons déjà vérifié l'hypothèse d'identifiabilité du modèle (pour l'obtention des solutions) lorsque nous avons présenté le VIF pour la détection d'éventuelles colinéarités entre les variables. Mais nous devons également vérifier les hypothèses rappelées ci-dessous :

Hypothèses Modèle Gaussien

Nous formulons les hypothèses suivantes pour notre modèle linéaire gaussien :

1. $(Y_i, X_i)_{i=1}^n$ doivent être *i.i.d.*, *i.e.* indépendantes et identiquement distribuées,
2. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$
3. $\varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$: hypothèse d'homoscédasticité.

Il s'agit donc essentiellement de vérifier les hypothèses relatives aux **résidus** $\hat{\varepsilon}_i$ du modèle.

Analyse des résidus On rappelle que les résidus sont définis, pour tout entier $i \in \llbracket 1, n \rrbracket$ par

$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

La validation des hypothèses se fera essentiellement à l'aide d'un graphique. Plus précisément, on représentera graphiquement les résidus *normalisés*.

On rappelle également que l'on a

$$\text{Var}[\hat{\varepsilon}] = \sigma^2(\mathbf{I} - \mathbf{H}),$$

où \mathbf{H} est la matrice de projection orthogonale sur l'espace engendré par \mathbf{X} . en particulier, nous avons donc

$$\text{Var}[\hat{\varepsilon}] = \sigma^2(\mathbf{I} - \mathbf{H}_{i,i}),$$

A nouveau, nous devons utiliser une estimation $\hat{\sigma}^2$ de σ^2 pour effectuer notre analyse. Soit

$$\text{Var}[\hat{\varepsilon}] = \hat{\sigma}^2(\mathbf{I} - \mathbf{H}_{i,i}).$$

Les résidus n'ont pas la même variance potentiellement (cela dépend de la valeur $H_{i,i}$, on va donc les normaliser :

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma} \sqrt{1 - H_{i,i}}}.$$

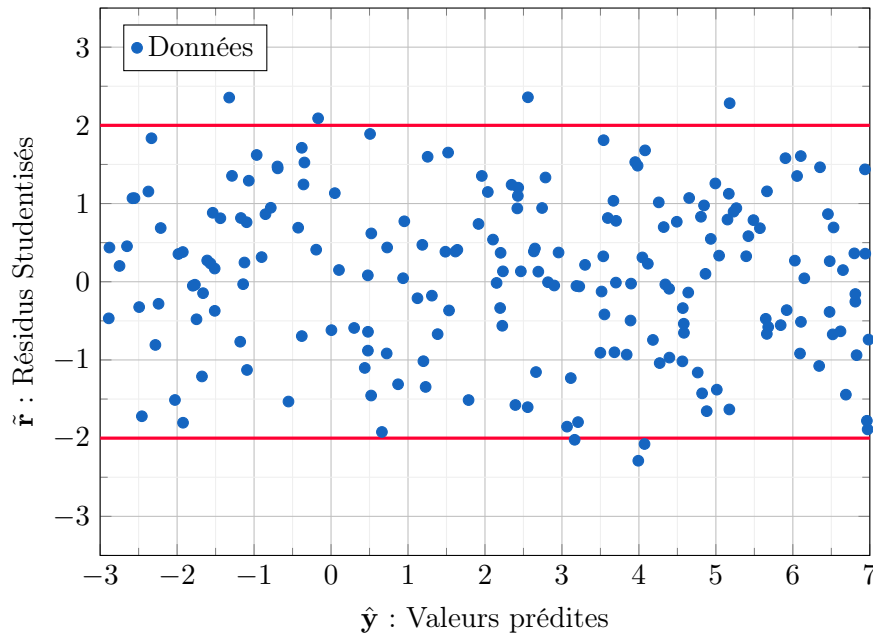
Ces résidus r_i ainsi obtenus sont appelés des **résidus standardisés** où $\hat{\sigma}^2 = \frac{1}{n - (p + 1)} \sum_{i=1}^n \hat{\varepsilon}_i^2$.

Il y a cependant un problème dans la définition de ces résidus, c'est que la valeur $\hat{\varepsilon}_i$ apparaît aussi bien au dénominateur qu'au numérateur, rendant ces deux quantités dépendantes, cela peut nuire à l'analyse de l'hypothèse *d'homoscedasticité*.

On va donc considéré des résidus dits **studentisés** $\tilde{r}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)} \sqrt{1 - H_{i,i}}}$,

où $\hat{\sigma}_{(i)} = \frac{1}{n - p - 2} \sum_{j \neq i}^n \hat{\varepsilon}_j^2$.

On va donc effectuer un graphe des **résidus studentisés en fonction des valeurs prédites**. Ce choix s'explique par le Théorème de Cochran qui garantit que les valeurs prédites \hat{y}_i et les résidus associés $\hat{\varepsilon}_i$ sont indépendantes. Il est donc à privilégier par rapport aux graphes $(\hat{\varepsilon}_i, X_i)$ ou encore $(\hat{\varepsilon}_i, Y_i)$.



Pour de grandes valeurs de n , on sait que 95% des valeurs de la loi de Student doivent se trouver dans l'intervalle $[-2, 2]$. Si trop de valeurs se trouvent en dehors de cet intervalle, on ne pourra pas dire que σ^2 est indépendant de X_i , contredisant ainsi l'hypothèse d'homoscédasticité.

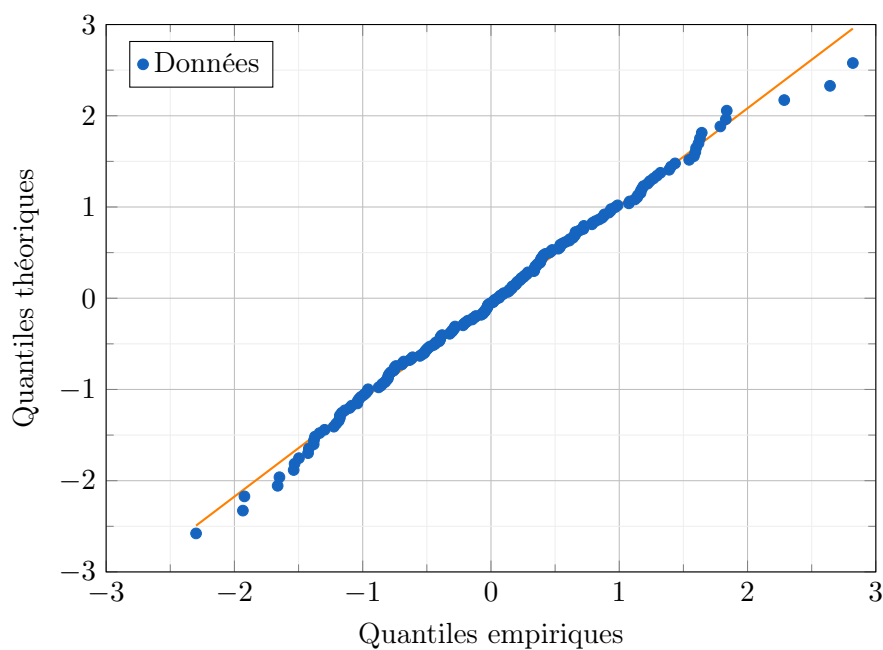
On va maintenant regarder comment tester la **normalité des résidus**. Nous pouvons le faire de deux façons : (i) à l'aide d'un test statistique ou (ii) à l'aide d'une méthode graphique.

Pour la première approche (i) consiste à effectuer un test de *Shapiro Wilk*. Le test prend la forme suivante :

$$H_0 : \text{les résidus sont normalement distribués v.s. } H_1 : \text{ils ne le sont pas}$$

Ce test est extrêmement puissant, mais aussi extrêmement rigide ce qui le rend peu utile en pratique car il aura tendance à rejeter très souvent l'hypothèse de normalité.

On préférera donc employer une méthode graphique (ii) qui repose sur la comparaison des quantiles empiriques des résidus aux quantiles théoriques de la loi normale centrée réduite.



On admettra que l'hypothèse de normalité ne sera pas contredite si les points sont globalement alignés.

Détection d'outliers Il faut parler de deux choses : (i) points leviers via les *hat values* et (ii) la distance de cook.

Parler des points leviers + hat values (technique de nettoyage du jeu de données en supprimant les points aberrants).

Cela finira la partie sur le modèle multiple.

6 Régression linéaire avec

A faire au plus vite pour illustrer les différentes notions et montrer, sur un exemple, comment construire un bon modèle.

Troisième partie

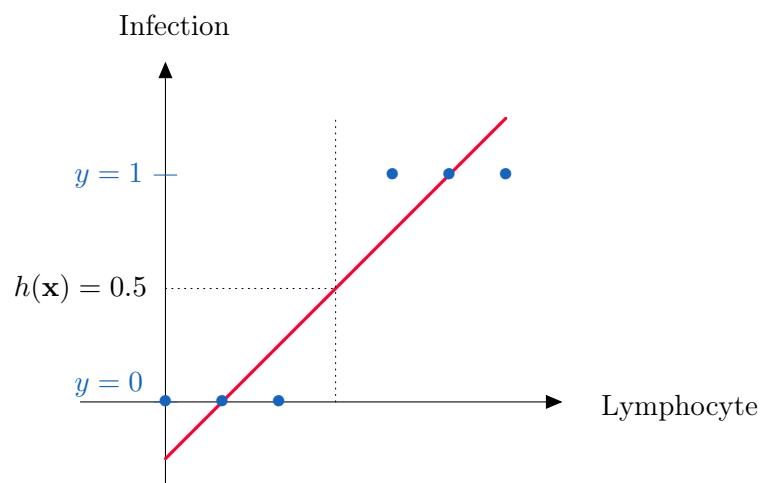
Modèles Linéaires Généralisés

6.1 Vers l'intérêt de la régression logistique

Plaçons dans un contexte un peu différent de celui que nous avons traité jusqu'à présent, et considérons l'exemple suivant.

On cherche à construire un modèle de régression capable de déterminer si un individu est atteint ou non d'une infection en fonction de sa numération en lymphocytes. La variable prédite peut prendre deux valeurs : 1 si la personne a une infection et 0 sinon.

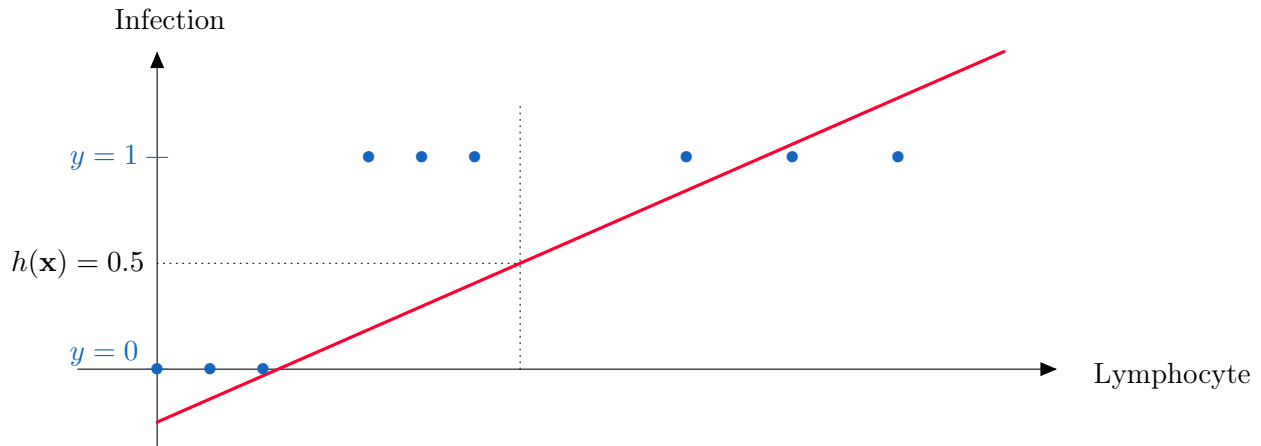
À première vue, rien ne nous empêche d'apprendre un modèle linéaire pour tenter d'ajuster notre nouveau nuage de points, comme illustré ci-dessous.



Il suffira alors de prendre un seuil, sur les valeurs prises par notre hypothèse h , au-delà duquel un individu sera considéré comme malade, *e.g.* on considère qu'un exemple \mathbf{x} appartient à la classe positive ($y = 1$) lorsque l'hypothèse h renvoie une valeur supérieure à 0.5 (*i.e.* négatif sur la partie gauche et positif sur la partie droite). Dans cet exemple, cela fonctionne bien.

Considérons maintenant un autre cas où le nombre de lymphocytes peut être extrêmement élevé, ce qui signifie que l'infection est grave. Ce nouvel ensemble de données

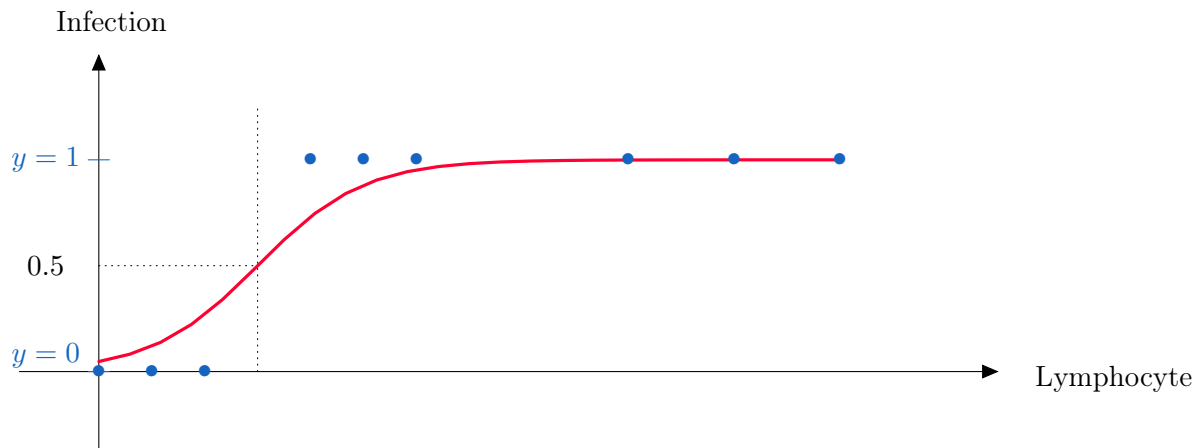
est représenté ci-dessous.



Cette fois-ci, nous constatons que si nous utilisons le même seuil, nous manquons des cas positifs ou des personnes infectées.

7 Vers la Régression Logistique

L'exemple précédent montre que la façon dont nous modélisons notre problème n'est pas bien choisie, nous avons besoin d'une structure différente, c'est-à-dire d'une courbe plus adaptée à la structure de nos données. Par exemple, nous avons besoin d'un modèle qui soit représenté comme suit :



Un tel modèle prend ses valeurs dans $[0, 1]$ et nous pouvons donc dire qu'il estime la probabilité d'avoir une infection. Pour transformer les valeurs prédites par un modèle de régression linéaire en probabilités, nous utilisons la fonction logistique, *i.e.* nous calculons :

$$\frac{1}{1 + \exp(-\mathbf{x}\beta)}.$$

On parle alors de *Régression Logistique*.

Régression Logistique : théorie et apprentissage Le modèle de *Régression Logistique*, également appelé modèle *logit*, a été introduit au milieu du 20^{me} siècle, mais l'utilisation des modèles *logit* remonte à la fin du 19^{me} siècle.

Pour estimer la probabilité qu'un exemple appartienne à une classe donnée, par exemple la classe positive : $\eta = Pr(Y = 1 | X)$, la régression logistique vise à calculer le logarithme du *odds*, c'est-à-dire le rapport des probabilités. Nous estimons ensuite le logarithme de ce rapport à l'aide d'un modèle linéaire :

$$\ln \left(\frac{Pr(y = 1 | \mathbf{x})}{Pr(y = 0 | \mathbf{x})} \right) = \mathbf{x}\beta + \varepsilon.$$

Ainsi, une fois les paramètres β du modèle sont appris, nous pouvons calculer la probabilité d'appartenir à la classe 1 :

$$Pr(y = 1 | \mathbf{x}) = \frac{\exp(\mathbf{x}\beta)}{1 + \exp(\mathbf{x}\beta)} = \frac{1}{1 + \exp(-\beta\mathbf{x})}.$$

Une telle fonction est appelée *fonction logistique* et prend ses valeurs dans $[0, 1]$. Un exemple \mathbf{x}_i est (généralement) prédit dans la classe 1 si $Pr(y = 1 | \mathbf{x}) > 0.5$, c'est-à-dire si $\mathbf{x}\beta > 0$. Compte tenu d'une tâche et d'un objectif, nous pouvons choisir de modifier ce seuil.

Pour estimer les paramètres du modèle, nous maximisons la vraisemblance des données $\mathcal{L}(\beta, S)$, où $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ est un ensemble de m exemples.

$$\begin{aligned} \mathcal{L}(\beta, S) &= \prod_{i=1}^m Pr(Y = y_i | X = \mathbf{x}_i), \\ &\downarrow \text{on sépare } y_i = 0 \text{ et } y_i = 1 \\ &= \prod_{i=1, y_i=1}^m Pr(Y = y_i | X = \mathbf{x}_i) \times \prod_{i=1, y_i=0}^m Pr(Y = y_i | X = \mathbf{x}_i), \\ &\downarrow \text{on utilise le fait que l'on suit une loi de Bernoulli} \\ &\downarrow \text{nous n'avons que deux issues possibles} \\ &= \prod_{i=1}^m \left(\frac{1}{1 + \exp(-\mathbf{x}_i\beta)} \right)^{y_i} \times \left(\frac{1}{1 + \exp(\mathbf{x}_i\beta)} \right)^{(1-y_i)}. \end{aligned}$$

Notez que nous préférons généralement minimiser la log-vraisemblance négative des données :

$$\begin{aligned}\ell(\boldsymbol{\beta}, S) &= -\ln(\mathcal{L}(\boldsymbol{\beta}, S)), \\ &= -\sum_{i=1}^m y_i \ln\left(\frac{1}{1 + \exp(-\mathbf{x}_i \boldsymbol{\beta})}\right) + (1 - y_i) \ln\left(1 - \frac{1}{1 + \exp(-\mathbf{x}_i \boldsymbol{\beta})}\right).\end{aligned}$$

Ce faisant, nous trouvons la fonction de perte logistique introduite précédemment. Dans ce qui suit, par souci de simplicité, nous fixerons $g(\boldsymbol{\beta}, \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x} \boldsymbol{\beta})}$. On est donc ramené à résoudre le problème d'optimisation suivant

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} -\frac{1}{m} \sum_{i=1}^m y_i \ln(g(\boldsymbol{\beta}, \mathbf{x}_i)) + (1 - y_i) \ln(1 - g(\boldsymbol{\beta}, \mathbf{x}_i)).$$

Nous divisons la perte par un facteur m afin d'être cohérent avec la notion de *moyenne des erreurs*

Par rapport au modèle linéaire de régression des moulinets, il n'existe pas de solutions analytiques. Cependant, le problème étant convexe, nous pouvons utiliser un algorithme basé sur le gradient pour trouver une solution.

Apprentissage du modèle On peut calculer le gradient de la fonction ℓ par rapport au vecteur $\boldsymbol{\beta}$. On a donc

$$\begin{aligned}\nabla \ell(\boldsymbol{\beta}, S) &= \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0}(\boldsymbol{\beta}, S) \\ \frac{\partial \ell}{\partial \beta_1}(\boldsymbol{\beta}, S) \\ \vdots \\ \frac{\partial \ell}{\partial \beta_d}(\boldsymbol{\beta}, S) \end{bmatrix}, \\ &= -\sum_{i=1}^m -y_i(1 - g(\boldsymbol{\beta}, \mathbf{x}_i))\mathbf{x}_i + (1 - y_i)g(\boldsymbol{\beta}, \mathbf{x}_i)\mathbf{x}_i, \\ &= -\sum_{i=1}^m \left(y_i - \frac{1}{1 + \exp(-\mathbf{x}_i \boldsymbol{\beta})}\right) \mathbf{x}_i.\end{aligned}$$

Nous pouvons ensuite appliquer l'algorithme de descente de gradient en utilisant l'expression ci-dessus du gradient de la log-vraisemblance négative (on l'appellera fonction de coût) :

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \eta \nabla \ell(\boldsymbol{\beta}^{(k)}, S),$$

$k = 1, 2, \dots$ et η est le pas d'apprentissage.

La plupart du temps, nous utilisons l'algorithme de descente de gradient de **Newton-Raphson** pour minimiser notre fonction de coût, *i.e.* nous utilisons la matrice hessienne de ℓ dans notre procédure de minimisation au lieu du pas d'apprentissage η .

Cette matrice hessienne est donnée par

$$\nabla^2 \ell(\boldsymbol{\beta}, S) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2}(\boldsymbol{\beta}, S) & \dots & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_d}(\boldsymbol{\beta}, S) \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0}(\boldsymbol{\beta}, S) & \dots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_d}(\boldsymbol{\beta}, S) \\ \vdots & & \vdots \\ \frac{\partial^2 \ell}{\partial \beta_d \partial \beta_0}(\boldsymbol{\beta}, S) & \dots & \frac{\partial^2 \ell}{\partial \beta_d^2}(\boldsymbol{\beta}, S) \end{bmatrix} = \sum_{i=1}^m g(\boldsymbol{\beta}, \mathbf{x}_i) (1 - g(\boldsymbol{\beta}, \mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^\top.$$

On peut exprimer la matrice hessienne sous une forme plus compacte comme suit :

$$\nabla^2 \ell(\boldsymbol{\beta}, \mathbf{X}) = \mathbf{X}^\top \mathbf{G} \mathbf{X},$$

où $\mathbf{X} \in \mathbb{R}^{m \times (d+1)}$ est la matrice de design (matrice *i.e.* des données) et $\mathbf{G} \in \mathbb{R}^{m \times m}$ est la matrice définie par :

$$G = \begin{bmatrix} g(\boldsymbol{\beta}, \mathbf{x}_1) (1 - g(\boldsymbol{\beta}, \mathbf{x}_1)) & 0 & \dots & \dots & 0 \\ 0 & g(\boldsymbol{\beta}, \mathbf{x}_2) (1 - g(\boldsymbol{\beta}, \mathbf{x}_2)) & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & g(\boldsymbol{\beta}, \mathbf{x}_m) (1 - g(\boldsymbol{\beta}, \mathbf{x}_m)) \end{bmatrix}.$$

Notez qu'avec l'expression ci-dessus, la matrice hessienne est exprimée comme une combinaison linéaire positive de matrices de Gram et est donc une matrice semi-définie positive. L'algorithme de Newton-Raphson est donc le suivant :

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\nabla^2 \ell(\boldsymbol{\beta}^{(k)}, S) \right)^{-1} \nabla \ell(\boldsymbol{\beta}^{(k)}, S),$$

qui présente un taux de convergence plus rapide que l'algorithme standard de descente de gradient à pas constant.

8 Régression à noyaux

9 Modèles (à effets) Mixtes

Références

- [Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6) :716–723.
- [Gilmour, 1996] Gilmour, S. G. (1996). The interpretation of mallows’s cp-statistic. *Journal of the Royal Statistical Society Series D : The Statistician*, 45(1) :49–56.
- [Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461 – 464.