# Vraisemblance et Moindres Carrés

Projet de Calcul Numérique en GIS3

Janvier 2019

En statistiques (et donc en data sciences), le concept de *vraisemblance* est très important et de nombreux problèmes se résolvent en maximisant une vraisemblance, qui dépend du problème. Déterminer les valeurs de paramètres qui maximisent un critère (la vraisemblance est un exemple de critère) c'est résoudre un *problème d'optimisation*. Parmi toutes les méthodes d'optimisation disponibles, la méthode des *moindres carrés* est l'une des plus connues et des plus utilisées. Il se trouve que dans certains cas, la vraisemblance peut se maximiser par moindres carrés alors que dans d'autres cas, les moindres carrés ne s'appliquent pas.

Ce document apporte un éclairage sur ces questions. Plusieurs cours de GIS s'y rattachent très naturellement : *calcul numérique*, *régression linéaire*, *statistiques inférentielles*, *probabilités*, *classification supervisée*. Voir également [2, chapter 8].

## Table des matières

# 1 Présentation — Statistiques Inférentielles

On se donne $n$ points $(x_i, y_i) \in \mathbb{R}^2$. On interprète les $n$ ordonnées $y_i$ comme les réalisations de $n$ variables aléatoires indépendantes $Y_i$. On suppose ici qu'elles vérifient le modèle statistique suivant (on raisonne sur un modèle particulier pour simplifier mais le raisonnement se généralise) :

$$Y_i \;=\; \theta_2\, x_i^2 + \theta_1\, x_i + \theta_0 + E_i\,.$$

Les variables aléatoires $E_i$ du modèle représentent les aléas. Les densités de probabilité $f_{i,\theta}$ des $Y_i$ dépendent de trois paramètres $\theta_2$, $\theta_1$ et $\theta_0$.

On rappelle que si $\varepsilon > 0$ est petit, le produit $2\,\varepsilon\, f_{i,\theta}(y)$ — où $y$ est une variable — est une approximation de la probabilité que $Y_i \in [y-\varepsilon, y+\varepsilon]$. Cette densité $f_{i,\theta}$, évaluée en la valeur observée $y_i$, devient une fonction des paramètres $\theta_2, \theta_1, \theta_0$ et à image dans $\mathbb{R}$, qu'on appelle la *vraisemblance correspondant à l'observation de la variable aléatoire* $Y_i$. Parce qu'on a supposé les $Y_i$ [ ], la vraisemblance correspondant à l'observation des $n$ variables aléatoires $Y_1, \ldots, Y_n$ est égale au produit des vraisemblances correspondant à l'observation de chacun des $Y_i$. On la note souvent $\mathscr{L}$ (pour *likelihood*) :

$$\mathscr{L}(\theta; y_1, \ldots, y_n) \;=\; \prod_{i=1}^{n} f_{i,\theta}(y_i)\,.$$

Le problème posé consiste à déterminer la valeur du vecteur de paramètres $\theta = (\theta_2, \theta_1, \theta_0)$ pour laquelle cette vraisemblance est maximale.

Pour expliciter les formules, il faut faire quelques hypothèses supplémentaires : supposons que les variables aléatoires $E_i$ soient centrées et aient chacune une distribution de probabilité gaussienne avec une même (pour simplifier) variance $\sigma^2 > 0$. Alors les variables $Y_i$ ont aussi des distributions de probabilité gaussiennes et

$$\begin{aligned}\mathscr{L}(\theta; y_1, \ldots, y_n) \;&=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi\,\sigma^2}} e^{-\frac{(y_i-(\theta_2\, x_i^2+\theta_1\, x_i+\theta_0))^2}{2\,\sigma^2}}\,,\\ &=\; \left(\frac{1}{2\,\pi\,\sigma^2}\right)^{\frac{n}{2}}\; e^{-\frac{1}{2\,\sigma^2}\,\sum_{i=1}^{n}(y_i-(\theta_2\, x_i^2+\theta_1\, x_i+\theta_0))^2}\,.\end{aligned}$$

Parce que la fonction logarithme est [ ], maximiser $\mathscr{L}$, c'est maximiser son logarithme

$$\ln \mathscr{L}(\theta; y_1, \ldots, y_n) = \boxed{\phantom{\quad}}$$

C'est aussi minimiser son opposé. Le problème posé se ramène donc à déterminer les valeurs de $\theta_2$, $\theta_1$ et $\theta_0$ qui minimisent la somme de carrés

$$\mathscr{L}_2 = \sum_{i=1}^{n} (y_i - (\theta_2\, x_i^2 + \theta_1\, x_i + \theta_0))^2\,. \tag{1}$$

## 1.1 Cas de la Régression Linéaire

Dans ce cas, le modèle devient

$$Y_i = \boxed{\phantom{xxxxxxxxxx}}$$

Le logarithme de la vraisemblance s'écrit

$$\ln \mathscr{L}(\theta; y_1, \ldots, y_n) = \boxed{\phantom{xxxxxxxxxxxxxxxx}}$$

Maximiser ce logarithme revient à minimiser la somme de carrés

$$\mathscr{L}_2 = \boxed{\phantom{xxxxxxxxxx}}$$

## 1.2 Équation Matricielle

Avant d'aller plus loin, il est utile d'introduire l'équation matricielle suivante :

$$\underbrace{\begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \theta_2 \\ \theta_1 \\ \theta_0 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{b}\,.$$

Poser que les trois paramètres sont solution de l'équation matricielle $A\,x = b$, c'est poser que les $n$ points $(x_i, y_i)$ appartiennent au graphe défini par le modèle, pour ces valeurs des paramètres, c'est-à-dire à la parabole d'équation $y = \theta_2\, x^2 + \theta_1\, x + \theta_0$. Avec ces notations et en utilisant une norme vectorielle, la somme de carrés (1) à minimiser peut se reformuler en $\boxed{\phantom{xxxxxx}}$.

## 1.3 Exemple

Prenons

| $x_i$ | 1 | 3 | 4 | 7 |
|---|---|---|---|---|
| $y_i$ | 2 | $-1$ | 2 | 3 |

Le système $A\,x = b$ correspondant est

$$\boxed{\phantom{\begin{matrix}0\\0\\0\end{matrix}}} \begin{pmatrix} \theta_2 \\ \theta_1 \\ \theta_0 \end{pmatrix} = \boxed{\phantom{\begin{matrix}0\\0\\0\end{matrix}}}\,.$$

## 1.4 Recherche des Extrema de la Somme de Carrés

### 1.4.1 Principe

Pour déterminer les extrema locaux d'une fonction $y = f(x)$ d'une variable réelle, on commence par chercher les points du graphe de $f$ à tangente horizontale, c'est-à-dire les solutions de l'équation $f'(x) = 0$. Cette idée se généralise aux fonctions de plusieurs variables réelles. Ainsi, pour déterminer les extrema locaux de la fonction de deux variables réelles $z = f(x, y)$, on cherche les points du graphe de $f$ à plan tangent horizontal. Les coordonnées $x$ et $y$ de ces points s'obtiennent en résolvant le système d'équations formé par les deux dérivées partielles de $f$, c'est-à-dire

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

Cherchons par exemple les extrema locaux de la fonction de deux variables réelles $z = f(x, y)$ où

$$f(x, y) = x^2 + x\,y + y^2 - 3\,x - 6\,y\,.$$

Les dérivées partielles de $f$ sont :

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

Ces dérivées partielles forment un système d'équations ayant pour unique solution

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \boxed{\phantom{\begin{matrix}0\\0\end{matrix}}}\,.$$

La valeur de $z$ correspondante est $\bar{z} = f(\bar{x}, \bar{y}) = -9$. Le point $(\bar{x}, \bar{y}, \bar{z})$ est donc un extremum local du graphe de $f$ (c'est en fait un minimum).

### 1.4.2 Maximisation de la Vraisemblance

On cherche un minimum de la somme de carrés (1). La fonction $\mathscr{L}_2$ à minimiser est une fonction des trois paramètres du modèle : $\theta_2$, $\theta_1$ et $\theta_0$. On obtient le minimum de $\mathscr{L}_2$, c'est-à-dire le maximum de vraisemblance, en résolvant le système des dérivées partielles de $\mathscr{L}_2$ par rapport à ces variables. En d'autres termes, on pose

$$\frac{\partial\,\mathscr{L}_2}{\partial\,\theta_\ell} = 0 \quad (1 \leq \ell \leq 3)\,,$$

c'est-à-dire

$$-2 \sum_{i=1}^{n} (y_i - (\theta_2 \, x_i^2 + \theta_1 \, x_i + \theta_0)) \, x_i^\ell = 0 \quad (1 \leq \ell \leq 3) \, .$$

Le facteur $-2$ peut être supprimé. Avec les notations introduites en section 1.2, ce système s'écrit matriciellement

$$A^T \, A \, x = A^T \, b \, . \tag{2}$$

Historiquement, on l'appelle *système des équations normales*.

En appliquant la règle sur la transposée d'un produit de matrices, on voit que $(A^T \, A)^T =$ [ ] et donc que $A^T \, A$ est une matrice [ ].

Par définition, le rang d'une matrice de dimension $m \times n$ est le nombre maximal de [ ]. On peut montrer (on admet) que le rang d'une matrice est égal au rang de sa transposée et donc que le rang de $A$ est inférieur ou égal à [ ]. Dans le cas qui nous intéresse, on a $m > n$. Par conséquent, $A$ est de rang maximal si et seulement si son rang est égal à [ ].

On peut montrer (on admet) que si $A$ est de rang maximal, alors $A^T \, A$ est *définie positive*.

# 2 Résolution — Algèbre Linéaire Numérique

## 2.1 Résolution du Système des Équations Normales

On applique l'algorithme du Commandant Cholesky sur la matrice symétrique $A^T \, A$. L'algorithme termine sans erreur si et seulement si $A^T \, A$ est [ ]. Si c'est le cas, on obtient alors une matrice triangulaire inférieure $L$ telle que $A^T \, A = L \, L^T$. Le système (2) se réécrit donc en :

$$L \underbrace{L^T \, x}_{v} = A^T \, b \, .$$

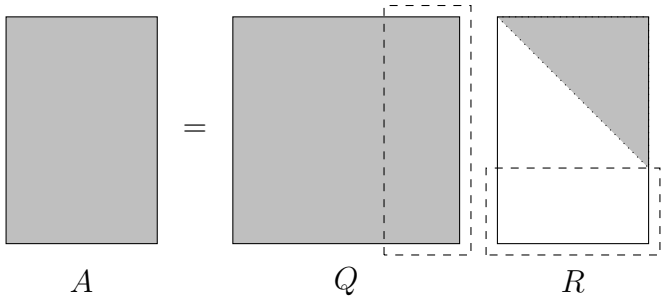
Il suffit alors d'enchaîner, dans l'ordre, les deux résolutions de systèmes triangulaires suivantes

[ ] .

Cette méthode est décrite dans [3, Algorithm 11.1, page 82]. Elle est prouvée numériquement instable [3, Theorem 19.3, page 142].

## 2.2 Résolution par la Factorisation QR

Le système des équations normales vient du fait qu'on a posé qu'un système de dérivées partielles s'annule. La méthode suivante s'appuie sur un autre type de raisonnement : l'existence de la factorisation $Q \, R$. On part du système $A \, x = b$ introduit en section 1.2. La matrice $A$ admet une factorisation $Q \, R$ de la forme suivante :




En partitionnant les matrices suivant les pointillés, on a donc :

$$A = Q\,R = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix}.$$

En multipliant les deux membres de l'équation par la transposée de $Q$ (qui est aussi son inverse, puisque $Q$ est [ ]), on obtient :

$$Q^T\,A\,x = Q^T\,Q\,R\,x = R\,x = Q^T\,b\,.$$

En faisant apparaître le partitionnement :

$$\begin{pmatrix} R_1 \\ 0 \end{pmatrix} x = \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} b\,.$$

Si on développe ce système d'équations bizarre, on observe qu'il comporte deux parties :

$$R_1\,x = Q_1^T\,b \qquad \text{et} \qquad 0 = Q_2^T\,b\,.$$

C'est un système sans solution (c'est normal, puisqu'il est équivalent à $A\,x = b$, qui n'en a pas non plus). On ne peut rien faire au sujet du système $0 = Q_2^T\,b$. Par conséquent, le vecteur $x$ qui minimise l'erreur $(\|b - A\,x\|_2)^2$ est la solution du système $R_1\,x = Q_1^T\,b$, qui se résout simplement puisqu'il s'agit d'un système [ ].

Cette méthode est décrite dans [3, Algorithm 11.2, page 83]. Pour peu qu'on calcule la factorisation $Q\,R$ par la technique des réflexions de [ ] ou une version stable de l'algorithme de Gram-Schmidt, cette méthode est numériquement *backward stable* (c'est [ ] que *stable*) [3, Theorem 19.1, page 140].

## 2.3 Résolution par la SVD

Cette méthode consiste à calculer la décomposition en valeurs singulières réduite $A = U\,\Sigma\,V^T$, résoudre le système diagonal $\Sigma\,v = U^T\,b$ puis prendre $x = V\,v$. Cette méthode est décrite dans [3, Algorithm 11.3, page 84]. Elle est commentée [3, pages 142-143]. Pour commencer, elle est *backward stable* [3, Theorem 19.4, page 143]. Et c'est même la seule méthode qui reste *stable* même dans les cas où la matrice $A$ n'est pas de rang maximal. Une des raisons tient au fait que les méthodes employées ont besoin de connaître le rang de $A$ et que la SVD fournit un des meilleurs algorithmes connus pour le calcul du rang.

# 3 Projet — Estimation de Paramètres pour la Fonction Logistique

L'importance de la fonction logistique (3) en dynamique de populations a été mise en évidence pour la première fois par Pierre-François Verhulst en 1838 [4].

$$y(x) \;=\; \frac{\kappa}{1+e^{\alpha-\rho\,x}}. \tag{3}$$

En dynamique de populations, la variable indépendante $x$ représenterait le temps, la variable dépendante $y(x)$ la population au temps $x$ et les trois lettres grecques $\kappa, \alpha, \rho$ des paramètres. La courbe de la fonction logistique a la forme d'une sigmoïde ayant pour asymptotes horizontales les droites $y = 0$ et $y = \kappa$.

Après les travaux de Verhulst, la fonction logistique a été redécouverte par Pearl en 1920 et a connu un immense succès pour la représentation de données expérimentales dans des domaines très variés. Voir par exemple [1, section 6.2, pages 148 et suivantes]. Une question très naturelle consiste alors, partant de données expérimentales, à déterminer les valeurs des trois paramètres $\kappa, \alpha, \rho$ qui font passer la sigmoïde au plus près de ces données.

On raisonne sur les données suivantes, adaptées de [1, Fig. 7.2, page 229]. À une constante additive près, il s'agit de $m = 10$ mesures de la quantité d'eau présente dans des œufs de *Locustana pardalina* en formation, à une température de 35 degrés.

| nb. jours $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| qté eau $y_i$ | .53 | .53 | 1.53 | 2.53 | 12.53 | 21.53 | 24.53 | 28.53 | 28.53 | 30.53 |

(4)

## 3.1 Estimation par moindres carrés linéaires

Introduisons la fonction logit définie par

$$\text{logit}(p) \;=\; \ln\left(\frac{p}{1-p}\right).$$

À partir de (3), on trouve

$$\frac{\kappa}{y} \;=\; 1+e^{\alpha-\rho\,x} \quad \text{et donc} \quad \text{logit}\left(\frac{\kappa}{y}\right) \;=\; \rho\,x-\alpha\,.$$

Supposons $\kappa$ connu. Il suffit alors d'appliquer, sur les ordonnées $y_i$ des points expérimentaux (tableau (4)), la transformation ci-dessus (logit) puis d'estimer les deux paramètres $\alpha$ et $\rho$ par la méthode des moindres carrés.

Comment déterminer $\kappa$ ? Visuellement, il est souvent facile d'estimer l'asymptote horizontale (d'équation $y = \kappa$) vers laquelle tend la sigmoïde. C'est la méthode préconisée dans les ouvrages anciens [1, page 150].

On remarque que l'emploi de cette méthode implique de choisir $\kappa > y_i$ pour $1 \leq i \leq m$ parce que logit$(p)$ n'est définie que pour $\boxed{\phantom{xxxxxx}}$. En prenant $\kappa = 30.54$ on trouve (voir Figure 1) :

$$\alpha = 5.163\,, \quad \rho = 1.188\,.$$

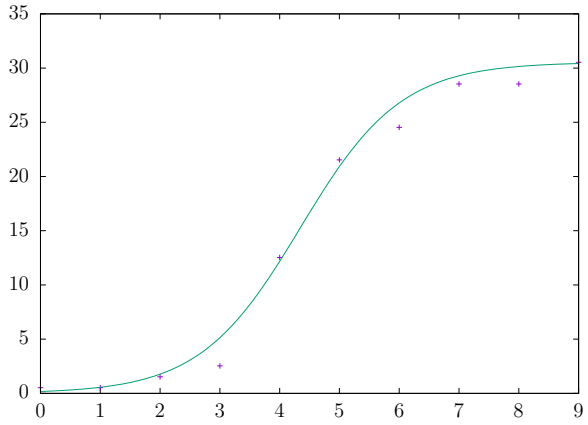



FIGURE 1 – Les données du tableau (4) et la courbe logistique obtenue pour $\kappa$, $\alpha$, $\rho$ = 30.54, 5.163, 1.188. La somme des carrés des écarts verticaux entre la courbe et les points vaut approximativement 3.968.

## 3.2 Estimation par moindres carrés non linéaires

L'estimation par la méthode de la section 3.1 n'est pas mauvaise mais n'est pas optimale (en général). En effet, il n'y aucune raison pour que $\kappa$ soit supérieur aux ordonnées de tous les points. C'est la raison pour laquelle les logiciels modernes utilisent une méthode plus sophistiquée qui est une variante de la méthode de Gauss-Newton et qu'on présente rapidement ici.

Soient $m$ le nombre de points expérimentaux et $n$ le nombre de paramètres à estimer. Sur notre exemple on a $m = 10$ et $n = 3$. L'idée consiste à construire une suite de vecteurs

$$v_0 = \begin{pmatrix} \kappa_0 \\ \alpha_0 \\ \rho_0 \end{pmatrix}\,, \quad v_1 = \begin{pmatrix} \kappa_1 \\ \alpha_1 \\ \rho_1 \end{pmatrix}\,, \quad v_2 = \begin{pmatrix} \kappa_2 \\ \alpha_2 \\ \rho_2 \end{pmatrix}\,, \quad \ldots$$

de $\mathbb{R}^n$ dont on espère qu'elle converge vers des valeurs qui minimisent la somme des carrés des écarts verticaux entre la courbe et les points. Pour cela, on introduit la fonction $s$ (pour *sigmoïde*)

$$s(\kappa, \alpha, \rho, x) \quad = \quad \frac{\kappa}{1 + e^{\alpha - \rho\, x}}$$

et on considère la fonction $r$ (pour *résidu*) de $\mathbb{R}^n$ dans $\mathbb{R}^m$ définie par

$$r(\kappa, \alpha, \rho) \quad = \quad \begin{pmatrix} s(\kappa, \alpha, \rho, x_1) - y_1 \\ s(\kappa, \alpha, \rho, x_2) - y_2 \\ \vdots \\ s(\kappa, \alpha, \rho, x_m) - y_m \end{pmatrix} .$$

Notons $J(\kappa, \alpha, \rho)$ la matrice jacobienne de la fonction $r$. Il s'agit de la matrice $m \times n$ définie par :

$$J(\kappa, \alpha, \rho) \quad = \quad \begin{pmatrix} s_{\kappa,1} & s_{\alpha,1} & s_{\rho,1} \\ s_{\kappa,2} & s_{\alpha,2} & s_{\rho,2} \\ \vdots & \vdots & \vdots \\ s_{\kappa,m} & s_{\alpha,m} & s_{\rho,m} \end{pmatrix} ,$$

où $s_{\kappa,i}$, $s_{\alpha,i}$ et $s_{\rho,i}$ désignent les dérivées partielles de $s$ par rapport à $\kappa$, $\alpha$ et $\rho$, évaluées en $x = x_i$. Sur l'exemple, cela donne :

$$s_{\kappa,i} = \frac{1}{1 + e^{\alpha - \rho\, x_i}} , \quad s_{\alpha,i} = -\kappa\, \frac{e^{\alpha - \rho\, x_i}}{(1 + e^{\alpha - \rho\, x_i})^2} , \quad s_{\rho,i} = \kappa\, x_i\, \frac{e^{\alpha - \rho\, x_i}}{(1 + e^{\alpha - \rho\, x_i})^2} .$$

Plaçons-nous à une itération $\ell$ et supposons $v_\ell$ déjà calculé. La méthode de Gauss-Newton consiste à construire le vecteur $r(\kappa_\ell, \alpha_\ell, \rho_\ell)$ et la matrice $J(\kappa_\ell, \alpha_\ell, \rho_\ell)$ — notés ci-dessous $r$ et $J$ pour simplifier — à résoudre le système d'équations linéaires

$$(J^T\, J)\, w \quad = \quad -J^T\, r \tag{5}$$

où $w$ est un vecteur de $n$ inconnues, puis à prendre

$$v_{\ell+1} \quad = \quad v_\ell + w\, .$$

La méthode de Gauss-Newton effectue autant de résolutions du système (5) qu'elle effectue d'itérations. Le système (5) est un système d'équations normales, comme dans la méthode historique de résolution des moindres carrés linéaires. Il peut se résoudre par la méthode du Commandant Cholesky. Bien que très utile, la méthode de Gauss-Newton n'est pas garantie (surtout sous cette forme-ci, qui est très simple) : elle peut ne pas converger du tout ou converger vers un vecteur différent du vecteur optimal. Le résultat peut en particulier varier en fonction du choix du vecteur initial $v_0$.

Dans le cas de la sigmoïde, on peut utiliser l'estimation obtenue par la méthode de la section 3.1 comme valeur initiale pour $v_0$. En appliquant cette idée sur l'estimation obtenue en fin de section 3.1, on trouve l'amélioration suivante (voir Figure 2) :

$$\kappa = 29.16\, , \quad \alpha = 5.8\, , \quad \rho = 1.3454\, .$$

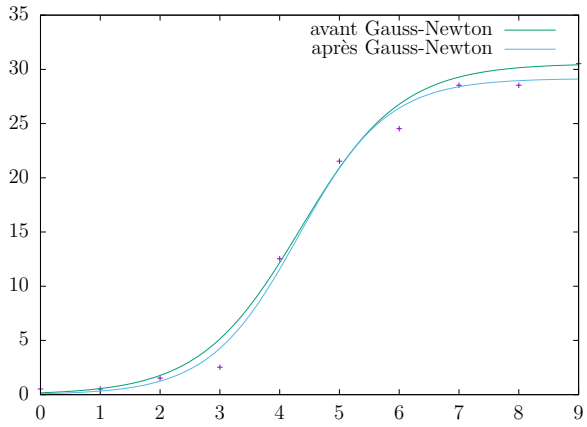



FIGURE 2 – Les données du tableau (4) et la courbe logistique obtenue pour $\kappa$, $\alpha$, $\rho = 29.16\,, 5.8\,, 1.3454$. L'asymptote horizontale est légèrement en-dessous du point d'ordonnée maximale. La somme des carrés des écarts verticaux entre la courbe et les points vaut approximativement 3.24.

## 3.3 Travail de programmation

Il s'agit d'écrire un programme FORTRAN qui lise les coordonnées des $m$ points formant une courbe sigmoïdale et qui calcule les valeurs des paramètres en appliquant les méthodes décrites dans ce sujet. Votre programme doit au minimum comporter une implantation de la méthode de la section 3.1. Une fois ce travail effectué, vous pouvez adapter la méthode pour estimer un quatrième paramètre ($\lambda$) en même temps que les trois autres, à partir du modèle

$$y(x) \;=\; \frac{\kappa}{1+e^{\alpha-\rho\,x}} + \lambda$$

et appliquer votre méthode sur les données réelles [1, page 229] :

| nb. jours $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| % eau $y_i$ | 51 | 51 | 52 | 53 | 63 | 72 | 75 | 79 | 79 | 81 |

(6)

**Question 1.** Quelle serait la matrice jacobienne avec le paramètre $\lambda$ ?

**Question 2.** (en option). Et si on remplaçait $\lambda$ par $\lambda + \mu$ ? Que deviendrait la matrice jacobienne ? Serait-elle de rang maximal ? La méthode des moindres carrés linéaires s'appliquerait-elle ?

### 3.3.1 Bibliothèques BLAS et LAPACK

Les bibliothèques `BLAS` permettent d'effectuer les calculs matriciels élémentaires :
— `IDAMAX` pour obtenir l'indice du maximum d'un vecteur, en valeur absolue,
— `DNRM2` pour le calcul de la norme 2 d'un vecteur,

— `DGEMM` pour le produit de matrices,
— `DGEMV` pour le produit matrice vecteur,
— `DTRSV` pour la résolution de système triangulaire.
La factorisation de Cholesky peut être calculée par la fonction `LAPACK` nommée `DPOTRF`.

# 4 Ouverture — Relation avec la Régression Logistique

La régression logistique est étudiée dans le cours de classification supervisée (GIS4). Comme dans le projet, il s'agit d'estimer les paramètres d'une fonction dont le graphe est une sigmoïde à partir de données expérimentales mais dans un cadre un peu différent.

Dans ce contexte-ci, les données expérimentales sont $n$ points $(x_i, y_i)$ où les abscisses $x_i \in \mathbb{R}$ mais où les ordonnées $y_i$ valent soit zéro soit un. On interprète les points $(x_i, y_i)$ comme les réalisations de $n$ couples indépendants de variables aléatoires $(X_i, Y_i)$ où les $X_i$ sont réelles et les $Y_i$ discrètes (binaires). On suppose que ces couples vérifient le modèle suivant, où $x$ est une variable réelle et $\alpha, \beta$ sont deux paramètres :

$$\mathrm{P}(Y_i = 1 \mid X_i = x) \;=\; \frac{e^{\alpha\, x+\beta}}{1+e^{\alpha\, x+\beta}} \quad \text{et donc} \quad \mathrm{P}(Y_i = 0 \mid X_i = x) \;=\; \frac{1}{1+e^{\alpha\, x+\beta}}.$$

Une astuce permet de combiner ces deux formules en une seule, où $x$ est une variable réelle et $y$ une variable binaire :

$$\mathrm{P}(Y_i = y \mid X_i = x) \;=\; \mathrm{P}(Y_i = 1 \mid X_i = x)^{y} \cdot \mathrm{P}(Y_i = 0 \mid X_i = x)^{1-y}\,.$$

Notons $f_i$ la densité de probabilité de $X_i$ et introduisons la fonction $\mathscr{D}_i$ suivante, qui peut être interprétée comme une densité, dans un sens généralisé à un couple de variables aléatoires, l'une continue, l'autre discrète :

$$\mathscr{D}_i(x,y) \;=\; \left(\frac{e^{\alpha\, x+\beta}}{1+e^{\alpha\, x+\beta}}\right)^{y} \cdot \left(\frac{1}{1+e^{\alpha\, x+\beta}}\right)^{1-y} \cdot f_i(x)\,.$$

On tient le même raisonnement qu'en section 1. Si $\varepsilon > 0$ est petit alors [ ] — où $x$ et $y$ sont des variables — est une [ ] que $(X_i, Y_i) \in [x-\varepsilon, x+\varepsilon] \times \{y\}$. Cette [ ], évaluée en l'observation $(x_i, y_i)$, devient une fonction [ ] et à image dans $\mathbb{R}$. Il s'agit de la vraisemblance correspondant à l'observation du [ ]. Elle s'écrit :

$$\mathscr{L}_i(\alpha, \beta; (x_i, y_i)) \;=\; \left(\frac{e^{\alpha\, x_i+\beta}}{1+e^{\alpha\, x_i+\beta}}\right)^{y_i} \cdot \left(\frac{1}{1+e^{\alpha\, x_i+\beta}}\right)^{1-y_i} \cdot f_i(x_i)\,.$$

La vraisemblance correspondant à l'observation des $n$ couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ est égale au produit :

$$\mathscr{L}(\alpha, \beta; (x_1, y_1), \ldots, (x_n, y_n)) \;=\; \prod_{i=1}^{n} \mathscr{L}_i(\alpha, \beta; (x_i, y_i))\,.$$

Le problème posé consiste à déterminer la valeur des paramètres $\alpha, \beta$ pour laquelle cette vraisemblance est maximale. On remarque qu'on atteint le même maximum 1) en ignorant le produit des $f(x_i)$ dans la formule ci-dessus, puisque les densités de probabilité $f_i$ ne dépendent pas des paramètres $\alpha, \beta$ ; 2) en maximisant le [ ] de la vraisemblance.

On aboutit ainsi à un problème [ ] qui n'est pas équivalent à un problème de moindres carrés — du moins tel qu'on l'a présenté dans ce document. Voir toutefois [2, section 4.4.1]. Ce type de problème peut se résoudre par une méthode connue sous le nom de *méthode de Newton* (ou *méthode de Newton-Raphson*) dont la méthode de Gauss-Newton est elle-même une variante, simplifiée pour le cas particulier des moindres carrés.

# 5 Consignes

Ce document contient un petit document de cours sur les relations entre vraisemblance et moindres carrés ainsi qu'un énoncé de projet de calcul numérique. Le document de cours contient des « trous » que vous êtes censés remplir au fil des cours du second semestre de GIS3. En fin de semestre, une évaluation d'une heure sans document sera organisée. Les questions seront liées aux contenus des « trous ». L'énoncé de projet de calcul numérique contient quelques questions ainsi qu'un travail de programmation. Le travail de programmation sera réalisé dans le cadre des séances de projet du cours de *calcul numérique*.

Le projet est à réaliser par binômes. Le rendu est constitué d'une archive à envoyer à `Francois.Boulier@polytech-lille.fr`. Cette archive doit contenir un rapport au format PDF, le ou les codes FORTRAN réalisés ainsi que les fichiers d'exemples.

La note porte essentiellement sur le rapport. Il doit être clair, comporter une introduction, une conclusion, une table des matières, vos réponses aux questions du projet et des instructions permettant de reproduire les figures présentes dans le rapport. Mettez en avant, dès l'introduction, les aspects positifs de votre travail. Le code doit compiler sans *warnings*, être raisonnablement indenté et commenté. L'archive doit être propre. Évitez les lettres accentuées et les espaces dans les noms de fichiers.

# Références


[1] L. C. Birch and H. G. Andrewartha. *The Distribution and Abundance of Animals.* The University of Chicago Press, 1954.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer Series in Statistics. Springer, 2nd edition, 2009. Available at `https://web.stanford.edu/~hastie/Papers/ESLII.pdf`.

[3] Lloyd Nicholas Trefethen and David Bau. *Numerical Linear Algebra.* SIAM, 1997.

[4] Pierre-François Verhulst. Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique*, 10 :113–121, 1838.