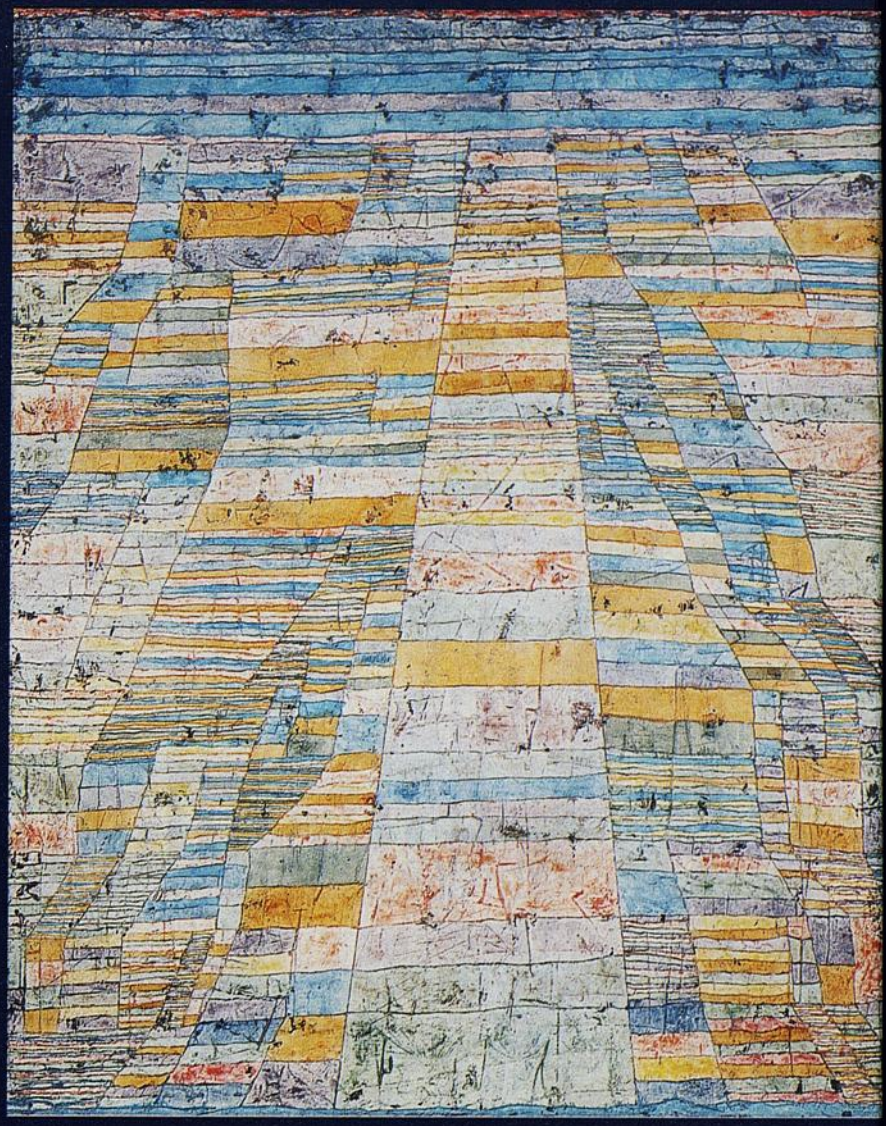
la régression non-linéaire

méthodes et applications en biologie

S. Huet
E. Jolivet
A. Messéan



MIEUX COMPRENDRE

INRA EDITIONS

la régression non-linéaire : méthodes et applications en biologie

la régression non-linéaire : méthodes et applications en biologie

Sylvie Huet, Emmanuel Jolivet, Antoine Messéan

INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE
147 rue de l'Université – 75338 Paris cedex 07

Mieux comprendre

Ouvrages parus dans la même collection :

Principes d'amélioration génétique des animaux domestiques
Francis MINVIELLE
1990, 211 p.

Cytogénétique des mammifères d'élevage
Paul C. POPESCU
1989, 114 p.

Les oligo-éléments en agriculture et élevage
Yves COÏC, Marcel COPPENET
1989, 114 p.

Eléments de virologie végétale
Pierre CORNUET
1987, 208 p.

L'épidémiologie en pathologie végétale
Mycoses aériennes
Frantz RAPILLY
1991, 317 p.

Amélioration des espèces végétales cultivées
Objectifs et critères de sélection
André GALLAIS, Hubert BANNEROT, éd.
1992, 748 p.

Photo de couverture
KLEE Paul, *Hauptweg und Nebenwege,* 1929
© Rheinisches Bildarchiv
Museum Ludwig Köln
© ADAGP, Paris, 1992

ISBN 2-7380-0413-X
ISSN 1144-7605

Avant-propos

Au début des années 70, alors que le cours des recherches en statistique appliquée n'entraînait pas naturellement le biométricien dans cette direction, Richard TOMASSONE, alors directeur du laboratoire de Biométrie du centre de recherche de Jouy-en-Josas, eut l'intuition de l'importance de la régression non-linéaire comme méthode d'analyse des données dans de nombreux domaines de la recherche agronomique. Avec Jean-Pierre VILA, il entreprit de rendre cette méthode accessible et populaire auprès des chercheurs de l'INRA. L'intérêt du laboratoire pour ce sujet, en permanence conforté par le renouvellement et l'importance pratique des applications qui en relèvent, ne s'est pas démenti depuis, même si l'évolution des disciplines a sensiblement modifié la façon de l'aborder.

Malgré cette démarche précoce, coïncidant pratiquement avec les premiers développements théoriques fondant la légitimité de la méthode des moindres carrés dans ce cadre, la régression non-linéaire demeure aujourd'hui une méthode un peu mystérieuse pour l'utilisateur. Plusieurs raisons pourraient expliquer cet état de fait. D'une part, la mise en œuvre des calculs réclame l'intervention directe de l'utilisateur, et les aspects algorithmiques, auxquels il est confronté en premier lieu, ont longtemps estompé la véritable nature statistique du problème. D'autre part, la théorie statistique sous-jacente n'est pas très simple, et tous les problèmes concrets qui relèvent de cette méthode ne sont pas encore résolus.

En écrivant cet ouvrage, notre objectif est donc de mettre à la disposition de l'utilisateur l'expérience de notre laboratoire, acquise tant par la multiplicité des ensembles de données traités que par les travaux de recherche théoriques qui y ont été conduits. En essayant de rendre accessibles les notions théoriques utiles à la compréhension de l'analyse, et en les illustrant par des exemples, nous espérons montrer qu'il n'y a en fait aucun mystère en la matière.

Comme chacun s'en doute, le passage de l'intention à la réalisation n'a pas été immédiat, et n'aurait sans doute jamais été accompli sans le soutien scientifique apporté par l'étroite coopération entretenue avec Olaf BUNKE et son équipe de l'Université Humboldt, à Berlin. Quant à notre *premier jet*, il a pu être modifié et amplement amélioré grâce à la lecture critique serrée de plusieurs collègues : Anestis ANTONIADIS, Sylvie AUDRAIN, Jacques BADIA, Jean-Pierre CAZES, Jean-Jacques CLAUSTRIAUX, François HOULLIER, Richard TOMASSONE et Bernard VAN CUTSEM. Nous les remercions tous bien vivement de leur précieuse collaboration.

Table des Matières

Liste des notations

Y_i est l'observation correspondant à la valeur x_i de la variable contrôlée : elles sont reliées par l'équation $Y_i = f(x_i, \theta) + \varepsilon_i$. L'indice i varie de 1 à n.

f est l'équation de régression.

ε_i est l'erreur associée à l'observation Y_i. C'est une variable aléatoire inobservable.

Y_{ij} apparaît dans le cas où plusieurs observations sont faites pour la même valeur de x_i. Alors, i varie de 1 à k et j varie de 1 à m. A l'observation Y_{ij} correspond l'erreur ε_{ij}.

θ est le vecteur des paramètres apparaissant dans la fonction de régression. Ses composantes sont notées $\theta_1, \ldots, \theta_a, \ldots, \theta_p$. Il varie dans une partie Θ de $\mathbf{R}^p$.

σ^2 est la variance (ou un facteur de proportionnalité de la variance) des erreurs ε_i.

β est un vecteur de paramètres intervenant éventuellement dans la définition de la variance de ε_i. Il est de dimension q.

$v(x_i, \theta, \beta)$ est la variance de l'observation Y_i (ou, ce qui revient au même, de l'erreur ε_i). Si cette variance ne dépend que du paramètre θ de la fonction de régression et du paramètre d'échelle σ^2, on la note $\sigma^2 v(x_i, \theta)$.

$\theta_\circ$, $\sigma^2_\circ$ et $\beta_\circ$ sont les *vraies valeurs* des paramètres.

$V_n(.)$ est la vraisemblance. $L_n(.)$ est égal à $-2n^{-1}\log V_n(.)$.

$Q_n(.)$ est la somme des carrés des écarts, $C_n = n^{-1}Q_n$.

$\widehat{\theta}_n$, $\widehat{\sigma}_n^2$ et $\widehat{\beta}_n$ sont les estimateurs (en général des moindres carrés ou du maximum de vraisemblance) de ces paramètres. D'autres notations sont employées pour d'autres estimateurs.

τ est le paramètre de $\mathbf{R}^{p+1}$ dont les p premières coordonnées sont celles de θ et dont la dernière coordonnée est σ^2, $\tau_\circ$ sa vraie valeur, $\widehat{\tau}_n$ son estimateur.

$\Gamma_{n\theta}$ est la matrice d'éléments $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{\partial f(x_i,\theta)}{\partial\theta_a}\frac{\partial f(x_i,\theta)}{\partial\theta_b}$. Elle intervient en particulier dans l'expression de l'estimation de la variance de $\widehat{\theta}_n$. Sa limite, quand n tend vers l'infini, est notée Γ_θ.

$I_n(\theta,\sigma^2)$ est la matrice d'information de Fisher. $I(\theta,\sigma^2)$ est la limite de $n^{-1}I_n(\theta,\sigma^2)$.

Y_i^*, $\widehat{\theta}_n^*$ et $\widehat{\sigma}_n^{*2}$ sont les versions bootstrap de Y_i, $\widehat{\theta}_n$ et $\widehat{\sigma}_n^2$. * désigne toujours un élément aléatoire correspondant à une expérience bootstrap.

$\widehat{\varepsilon}_i$ est le résidu $Y_i - f(x_i,\widehat{\theta}_n)$, $\widehat{e}_i$ est le résidu réduit (c'est-à-dire normalisé par une estimation de son écart-type) correspondant.

o(.) est une quantité telle que $\lim_{u\to 0}\mathrm{o}(u)/u = 0$. $\mathrm{o}(u)$ tend donc vers 0 plus vite que u; on dit que c'est un *infiniment petit* d'ordre inférieur à u.

1 Les modèles de régression non-linéaire

En statistique, le mot régression désigne un type de modèles bien précis. Or, la notion de modèle est absolument essentielle dans toute étude statistique. Il s'agit en effet de représenter une réalité de nature aléatoire et, au moyen de la confrontation de l'observation et du modèle, de déduire les lois, ou certains éléments des lois qui gouvernent cette réalité. Cette existence première du modèle est souvent oubliée par les utilisateurs de la statistique, qui ont alors tendance à réduire les méthodes qu'ils emploient à des recettes et à oublier les hypothèses sous-jacentes. Il est donc essentiel de les caractériser très précisément, avant de chercher à les exploiter. C'est pourquoi, dans cette première partie, nous allons définir les modèles de régression, et délimiter, au sein de cet ensemble, ceux que nous étudierons dans cet ouvrage. Deux éléments bien distincts constituent ces modèles : nous montrerons à l'aide d'exemples comment choisir a priori chacun d'eux.

1.1 Les modèles statistiques de régression

L'analyse des résultats d'une expérience relève de l'emploi d'un modèle de régression lorsque les observations qui en sont issues peuvent être représentées, chacune, comme la somme d'un terme systématique, dépendant de la valeur prise par une ou plusieurs autres variables, et de la réalisation d'une variable aléatoire. L'objectif est en général l'étude de la relation entre la variation d'une variable, par exemple le poids de matière sèche d'une talle de blé, et d'une ou plusieurs autres variables, par exemple la somme des températures moyennes journalières depuis la date de levée. Avant de revenir de manière plus précise sur cette définition, nous allons rappeler, à l'aide d'un exemple très simple, la notion de modèle statistique.

1.1.1 Pile ou face, quelques généralités

C'est le jeu de pile ou face qui nous fournit cet exemple. Si l'on suppose que l'on procède à n jets successifs d'une pièce de monnaie, indépendants entre eux, et tels que la probabilité d'obtenir pile soit la même à chaque jet, soit p, on sait bien que le nombre de fois où l'on obtient pile est une variable aléatoire N de loi binomiale $B(n,p)$, telle que

$$\Pr\{N=k\} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k},\ k=0,\ldots,n.$$

Dans un tel cadre, on peut supposer la connaissance de l'expérimentateur réduite à l'observation du nombre de pile et au fait que ce nombre suive une loi binomiale de paramètres n connu et p inconnu. Son souhait sera évidemment de connaître p aussi bien que possible, pour vérifier par exemple qu'il est égal à 0,5. Le modèle statistique de l'expérience, dont le résultat est ici réduit à N, est alors l'ensemble des lois binomiales $B(n,p)$, pour p variant de 0 à 1, soit

$$\mathcal{M} = \{B(n,p), p \in]0,1[\}\,.$$

Il convient a priori d'établir une distinction entre le paramètre du modèle, p, et la loi de l'observation, $B(n,p)$. Néanmoins, dans ce cas particulier, il y a identification entre ces deux notions. Nous verrons que ce n'est pas toujours le cas. D'une façon générale, le modèle statistique d'une expérience est un ensemble de lois de probabilités qui contient la loi de l'observation issue de cette expérience. Cet ensemble de lois de probabilité peut être décrit, complètement ou partiellement, par un ensemble de paramètres. L'objectif essentiel de la statistique est l'identification de la loi de probabilité de l'observation, ou, à défaut, des paramètres caractérisant cette loi, utiles à l'expérimentateur. Intuitivement, il est clair que le modèle doit être défini avec beaucoup de soin. S'il est trop grand, il risque d'être délicat d'y trouver la loi d'où est issue l'observation et impossible de l'exploiter. S'il est trop petit, ou mal défini, la loi que l'on cherche à identifier risque d'être en dehors de l'ensemble des lois de probabilité qui constituent le modèle. Nous allons préciser ces généralités dans le cadre des modèles de régression.

1.1.2 La régression

Si nous revenons à l'exemple de la croissance d'une talle de blé, il est raisonnable de penser qu'il existe une forte relation entre son poids de matière sèche à une date donnée et la somme de degrés×jours observée depuis la levée de la plante. Cette relation ne peut cependant pas être considérée comme déterministe, dans la mesure où il existe une variabilité dans la population des plantes étudiées, due à des facteurs génétiques et d'environnement, mais qui peut être considérée comme résiduelle et aléatoire vis-à-vis de la variation systématique induite par la température. Nous

représentons mathématiquement, et certainement de manière simplifiée, cette hypothèse en écrivant que, à une date où la somme de degrés×jours observée depuis la levée est x, le poids de matière sèche d'une talle prise au hasard dans la population est une variable aléatoire Y qui s'exprime comme la somme d'une quantité fonction de x, $f(x)$, inconnue, mais non aléatoire, et d'une variable aléatoire d'espérance nulle (puisqu'elle n'exprime qu'un écart résiduel à une tendance générale) ε, soit $Y = f(x) + \varepsilon$. Il résulte évidemment de cette dernière équation que l'espérance mathématique de Y est $f(x)$. L'équation $\mathbf{E}(Y) = f(x)$ est appelée équation de régression. En général Y est appelée la variable dépendante ou la réponse, x la variable indépendante ou la variable contrôlée et ε le résidu ou l'erreur.

Lorsqu'une expérience est destinée à étudier une relation entre deux quantités, le résultat est une série de couples de valeurs, ou plus exactement une série de valeurs de la variable dépendante obtenues pour des valeurs connues de la variable indépendante, soit $(y_1, x_1), \ldots, (y_n, x_n)$. Supposons que les quantités observées $y_1, \ldots, y_n$ sont une réalisation du vecteur aléatoire $Y_1, \ldots, Y_n$ dont chacune des composantes peut s'écrire sous la forme $Y_i = f(x_i) + \varepsilon_i$. Un modèle statistique de régression est donc un ensemble de lois de probabilité contenant celle de l'observation $Y_1, \ldots, Y_n$. Sa définition est équivalente à celle d'un ensemble de fonctions auquel appartient f et d'un ensemble de lois de probabilité auquel appartient la loi du vecteur $\varepsilon_1, \ldots, \varepsilon_n$. Même si c'est la fonction f qui intéresse essentiellement l'expérimentateur, chacune des deux composantes du modèle doit être définie avec beaucoup de soin.

La définition qui vient d'être donnée d'un modèle statistique de régression est très générale, beaucoup trop pour que l'on puisse proposer un seul type de méthodes pour analyser des données relevant de tels modèles. Il convient donc d'identifier, à l'intérieur de cet ensemble immense de modèles de régression possibles, des sous-ensembles relevant de types particuliers de traitements. Dans cette monographie, nous étudions l'un de ces sous-ensembles, dont nous précisons maintenant les limites.

1.1.3 La classe de modèles étudiée

Plusieurs restrictions sont imposées immédiatement pour définir le cadre de notre étude. Elles sont d'importances inégales. Les premières sont fondamentales : nous supposons que

- les variables observées, tant indépendantes que dépendantes, sont continues, c'est-à-dire qu'elles prennent leurs valeurs dans des intervalles de l'ensemble des nombres réels;
- les valeurs des variables indépendantes sont fixées;
- les diverses observations de la variable dépendante sont indépendantes entre elles.

En revanche, une restriction supplémentaire est adoptée uniquement par souci de simplicité : nous ne présentons explicitement ici que le cas où il existe une seule variable indépendante. Le cas où le modèle de régression s'écrit en fonction de plusieurs variables explicatives est une généralisation facile de ce cas particulier, seules les notations sont un peu plus compliquées. Nous présenterons un peu plus loin un exemple de modèle de régression non-linéaire avec deux variables indépendantes.

Ce cadre étant fixé, examinons le modèle de régression linéaire simple, connu et étudié depuis longtemps (voir par exemple TOMASSONE *et al.* [67]). Dans ce modèle, l'ensemble auquel appartient la fonction f est l'ensemble des fonctions linéaires de la forme $f(x) = \theta_1 + \theta_2 x$ où θ_1 et θ_2 sont deux nombres réels quelconques. Les erreurs sont des variables aléatoires indépendantes et de même loi, et cette loi est en outre souvent supposée gaussienne. Cette hypothèse détermine évidemment la loi du vecteur aléatoire $\varepsilon_1, \ldots, \varepsilon_n$. On peut donc représenter le modèle statistique de régression linéaire sous la forme

$$\mathcal{M} = \{f : f(x) = \theta_1 + \theta_2 x, \mathcal{G}\}$$

où $\mathcal{G}$ est l'ensemble des lois gaussiennes centrées sur $\mathbf{R}$. Une variable aléatoire gaussienne centrée étant complètement définie par sa variance, la donnée d'un élément g de $\mathcal{G}$ est équivalente à la donnée d'un nombre réel positif σ^2 si l'on suppose alors que g est une loi de probabilité de densité $\phi(x) = \left(\sqrt{2\pi}\sigma\right)^{-1} \exp\left(-x^2/2\sigma^2\right)$, $x \in \mathbf{R}$. Un élément de $\mathcal{M}$ est donc complètement déterminé par la donnée des trois nombres θ_1, θ_2 et σ^2 : les paramètres du modèle.