M1 Iséfar

Modèles de Régression
Chapitre 1 : Introduction

Cécile Durot
cecile.durot@gmail.com

Université Paris-Ouest-Nanterre-La Défense

2014-2015

Objectif

On souhaite "expliquer" une variable y (variable réponse ou endogène) par une variable x (variable explicative ou exogène) :

$$y \approx f(x),$$

i.e. comprendre comment évolue typiquement y lorsque x varie.

- x peut prendre différentes formes : qualitative, quantitative, vecteur de variables qualitatives et/ou quantitatives.
- On considèrera uniquement des cas où y est à valeurs réelles. Dans le cas d'une variable réponse qualitative, on codera donc les valeurs possibles de cette variable par exemple par 0,1,...

Attention :

- Expliquer ne signifie pas qu'il existe une relation de cause à effet.
- Le choix de la méthode dépend de la nature de y (qualitative ou quantitative).

Définition :

Disposant de n observations (x_i, y_i), $i = 1, \ldots, n$, poser un modèle de régression avec plan d'expérience déterministe (ou fix design) consiste à :

- supposer que pour tout i, x_i est déterministe tandis que y_i est la réalisation d'une variable aléatoire réelle Y_i satisfaisant

$$E(Y_i) = f(x_i)$$

  pour une fonction f inconnue,
- supposer que les variables Y_i sont indépendantes entre elles,
- supposer que $f \in \mathcal{F}$ où $\mathcal{F}$ est une famille donnée de fonctions,
- formuler éventuellement des hypothèses supplémentaires sur les lois de $Y_1, \ldots, Y_n$.

Ecriture des modèles de régression

Supposer $E(Y_i) = f(x_i)$ revient à supposer

$$Y_i = f(x_i) + \varepsilon_i \text{ pour tout } i = 1, \ldots, n$$

où les ε_i sont des variables aléatoires réelles centrées non observées. Il est équivalent de supposer les variables $Y_1, \ldots, Y_n$ indépendantes entre elles et de supposer les variables $\varepsilon_1, \ldots, \varepsilon_n$ indépendantes entre elles. Formuler des hypothèses sur les lois de $Y_1, \ldots, Y_n$ revient à formuler des hypothèses sur les lois de $\varepsilon_1, \ldots, \varepsilon_n$.

Attention : On suppose dans certains cas que les variables $\varepsilon_1, \ldots, \varepsilon_n$ sont i.i.d. Les variables $Y_1, \ldots, Y_n$ quant à elles sont indépendantes mais n'ont pas la même loi (pas la même espérance).

Ajustement

On souhaite ajuster le modèle, c'est-à-dire construire un estimateur $\widehat{f}$ de f dans le but

- de réaliser des prédictions $\widehat{Y} = \widehat{f}(x)$,
- ou d'inférer, c'est-à-dire de comprendre comment évolue Y lorsque x évolue (quelles sont les variables significatives, quelle est la relation entre Y et chacune des variables explicatives ...).

Dans le premier cas, $\widehat{f}$ peut être traité comme une "boite noire" au sens où l'on ne s'intéresse pas à sa forme exacte pourvu qu'il produise de bonnes prédictions. Dans le second cas, la forme de $\widehat{f}$ est importante.

Les modèles linéaires ou linéaires généralisés permettent une inférence relativement simple et interprêtable, mais peuvent fournir des prédictions moins précises que d'autres méthodes plus complexes.

Classification supervisée

Dans le cas d'une variable réponse y qualitative, l'objectif pourra être de bâtir une règle de classification : par exemple, dans le cas $y \in \{0, 1\}$ et pour une nouvelle variable explicative x_0, la règle de Bayes consiste à attribuer la classe

$$\begin{cases} 1 \text{ si } \widehat{f}(x_0) \geqslant 1/2 \\ 0 \text{ sinon} \end{cases}$$

à une future observation associée à x_0.

Modèles paramétriques ou non paramétriques

- Un modèle de régression est dit paramétrique si $\mathcal{F}$ peut s'écrire sous la forme

$$\mathcal{F} = \{f_\beta, \beta \in \mathcal{B}\}$$

où $\mathcal{B} \subset \mathbb{R}^p$ est connu et f_β est connue pour tout $\beta \in \mathcal{B}$. Cela signifie qu'il existe $\beta^\star \in \mathcal{B}$ inconnu tel que $f = f_{\beta^\star}$, c'est-à-dire que f est connue à une nombre fini de paramètres près. Un estimateur de f prend alors la forme $\widehat{f} = f_{\widehat{\beta}}$ où $\widehat{\beta}$ est un estimateur de β.

- Dans le cas contraire, le modèle est dit non paramétrique.

Modèles paramétriques linéaires ou non linéaires

Supposons que $f \in \mathcal{F}$ où

$$\mathcal{F} = \{f_\beta, \beta \in \mathcal{B}\}$$

avec $\mathcal{B} \subset \mathbb{R}^p$ connu et f_β connue pour tout $\beta \in \mathcal{B}$.

- Un tel modèle est dit linéaire si pour tout x, l'application

$$\begin{array}{ccc} \mathcal{B} & \to & \mathbb{R} \\ \beta & \mapsto & f_\beta(x) \end{array}$$

est linéaire, i.e., si pour tous $\beta, \beta' \in \mathcal{B}$ et tous $b, b' \in \mathbb{R}$,

$$f_{b\beta + b'\beta'}(x) = b f_\beta(x) + b' f_{\beta'}(x), \ \forall x,$$

i.e. si on peut construire un vecteur ligne $v(x)$ ne dépendant que de x tel que pour tout β,

$$f_\beta(x) = v(x)\beta, \ \forall x.$$

- Dans le cas contraire, le modèle est dit non linéaire.

Modèles de régression additifs

Dans le cas où $x = (x^{(1)}, \ldots, x^{(k)})$ est consitué de k variables explicatives, $k \geqslant 2$ entier, on a souvent recours aux modèles additifs, qui supposent

$$f(x) = \sum_{j=1}^{k} f_j(x^{(j)})$$

pour des fonctions $f_1, \ldots, f_k$ inconnues.

Exemple 1 : Etude de mortalité

- y : taux de mortalité
- x : dureté de l'eau

mesurés dans 61 grandes villes d'Angleterre et du Pays de Galles entre 1958 et 1964.

Compromis biais/variance

La qualité de l'ajustement peut se mesurer avec l'erreur quadratique moyenne :

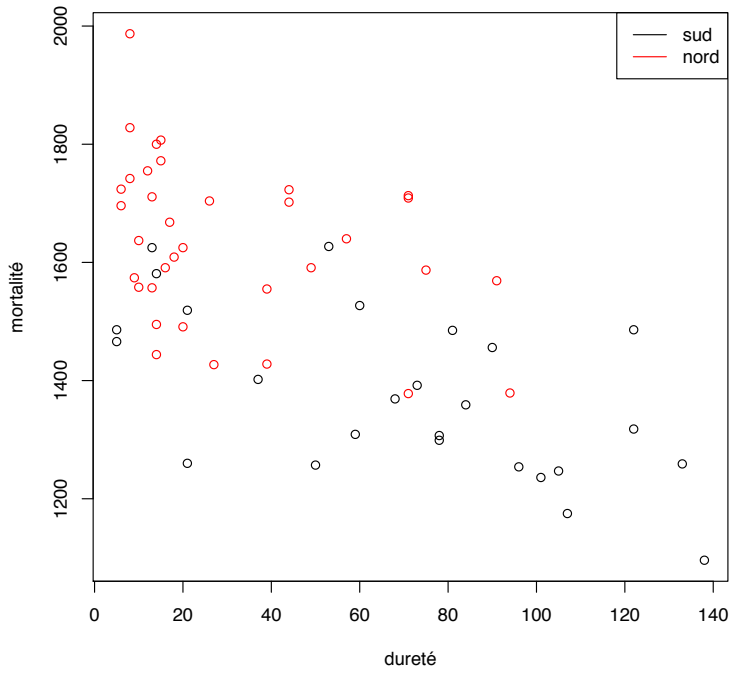
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2.$$

Modèle flexible $\Longrightarrow$ MSE faible. Cependant, un estimateur de très faible MSE sera typiquement très variable (sur-ajustement). Pour juger de la pertinence d'un modèle, on peut calculer l'estimateur sur un échantillon dit d'apprentissage, puis calculer l'erreur quadratique moyenne d'un échantillon test. Sur une donnée test, on a:

$$E(Y_0 - \widehat{f}(x_0))^2 = var(\widehat{f}(x_0)) + (\text{biais}(\widehat{f}(x_0)))^2 + var(\epsilon).$$

Le risque systématique est $var(\epsilon)$. On cherche un modèle (ou un estimateur) réalisant un bon compromis entre $var(\widehat{f}(x_0))$ et $(\text{biais}(\widehat{f}(x_0)))^2$ (le premier augmente, le second diminue lorsque la flexibilité augmente).

représentation de la mortalité en fonction de la dureté



Modèle pour les données Sud :

$$Y_i = \mu + ax_i + \varepsilon_i, \ i = 1, \ldots, n$$

où les ε_i sont i.i.d. gaussiennes centrées (modèle de régression linéaire simple : cadre paramétrique linéaire).

Droite ajustée : il s'agit de la droite d'équation

$$y = \widehat{\mu} + \widehat{a}x,$$

où $\widehat{\mu}$ et $\widehat{a}$ sont des estimateurs de μ et a.

Définition Exemples Démarche Plan

Modèle pour les données Sud :

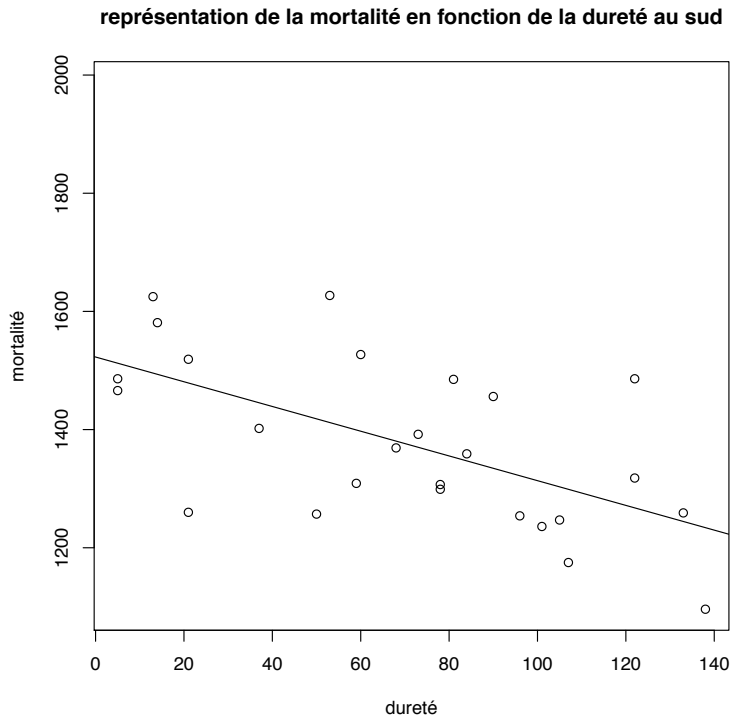
$$Y_i = \mu + ax_i + \varepsilon_i, \ i = 1, \ldots, n$$

où les ε_i sont i.i.d. gaussiennes centrées (modèle de régression linéaire simple : cadre paramétrique linéaire).

Droite ajustée : il s'agit de la droite d'équation

$$y = \widehat{\mu} + \widehat{a}x,$$

où $\widehat{\mu}$ et $\widehat{a}$ sont des estimateurs de μ et a.



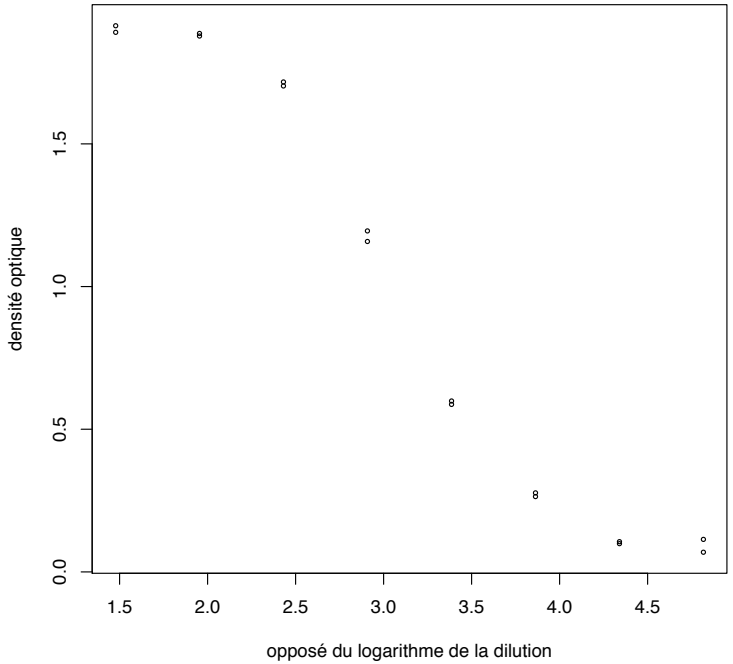
Exemple 2 : Essai Elisa

Pour mesurer le niveau d'anticorps d'un sérum de vache, on réalise diverses dilutions du sérum puis on mesure la densité optique (log décimal de l'opacité) pour chacune de ces dilutions.

- y : densité optique
- x : -log(dilution)

mesurés pour 8 dilutions différentes, avec une répétition pour chaque dilution.

Essai Elisa : représentation des données



Modèlisation par une courbe logistique :

$$Y_i = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(\beta_3(x_i - \beta_4)\right)} + \varepsilon_i, \ i = 1, \ldots, n$$

où les ε_i sont i.i.d. gaussiennes centrées et les β_j sont des réels inconnus. Cadre paramétrique non linéaire.

Courbe ajustée : il s'agit de la courbe d'équation

$$y = \widehat{\beta}_2 + \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{1 + \exp\left(\widehat{\beta}_3(x - \widehat{\beta}_4)\right)},$$

où $\widehat{\beta}_1, \ldots, \widehat{\beta}_4$ sont des estimateurs de $\beta_1, \ldots, \beta_4$.

Modèlisation par une courbe logistique :

$$Y_i = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(\beta_3(x_i - \beta_4)\right)} + \varepsilon_i, \ i = 1, \ldots, n$$

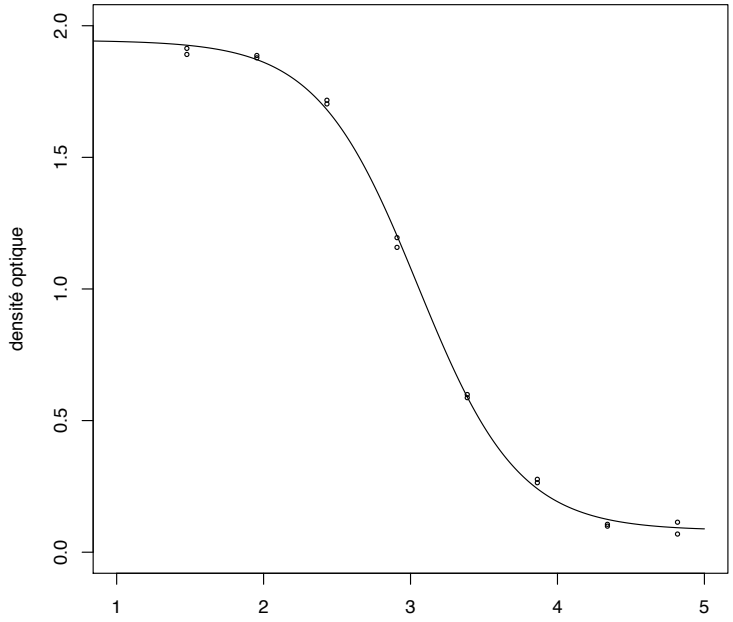
où les ε_i sont i.i.d. gaussiennes centrées et les β_j sont des réels inconnus. Cadre paramétrique non linéaire.

Courbe ajustée : il s'agit de la courbe d'équation

$$y = \widehat{\beta}_2 + \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{1 + \exp\left(\widehat{\beta}_3(x - \widehat{\beta}_4)\right)},$$

où $\widehat{\beta}_1, \ldots, \widehat{\beta}_4$ sont des estimateurs de $\beta_1, \ldots, \beta_4$.

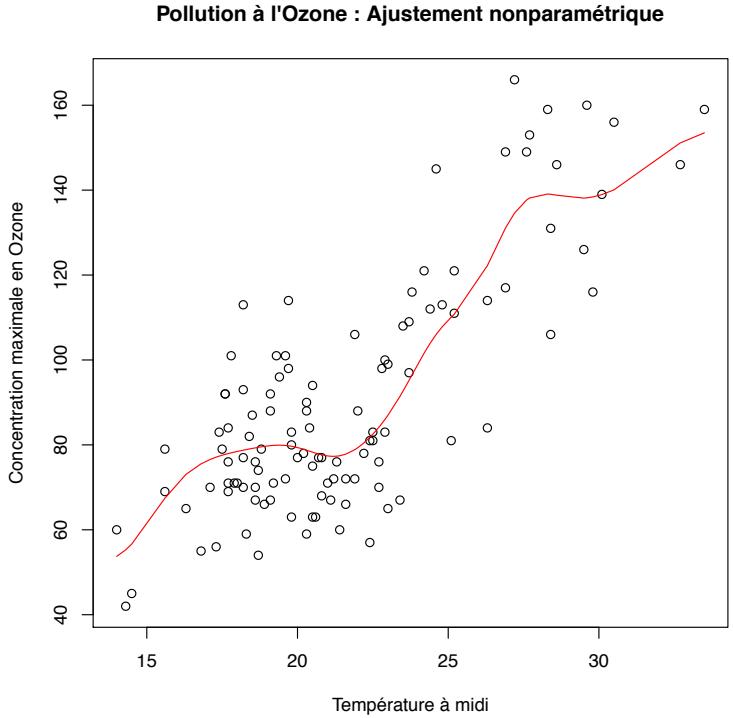
Essai Elisa : ajustement nonlinéaire logistique



Exemple 3 : Pollution à l'ozone

- y : concentration maximale en ozone
- x : température à midi

mesurés en un lieu donné et une journée donnée pendant n jours.
Ajustement non paramétrique.



Exemple 4 : Plasma

- $y = \begin{cases} 1 \text{ si l'individu est sain} \\ 0 \text{ sinon} \end{cases}$,
  au sens ESR (eurythorocyte sedimentation rate)
- $x =$ (mesure de γ-globulin,mesure de fibrinogène)

mesurés sur n individus.

Remarque : La loi de Y_i est entièrement déterminée par

$$f(x_i) = E(Y_i|x_i) = P(Y_i = 1|x_i).$$

Définition Exemples Démarche Plan

Exemple 4 : Plasma

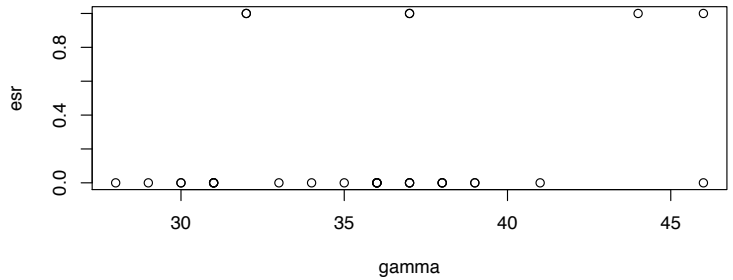
- $y = \begin{cases} 1 \text{ si l'individu est sain} \\ 0 \text{ sinon} \end{cases}$,
  au sens ESR (eurythorocyte sedimentation rate)
- $x =$ (mesure de γ-globulin,mesure de fibrinogène)

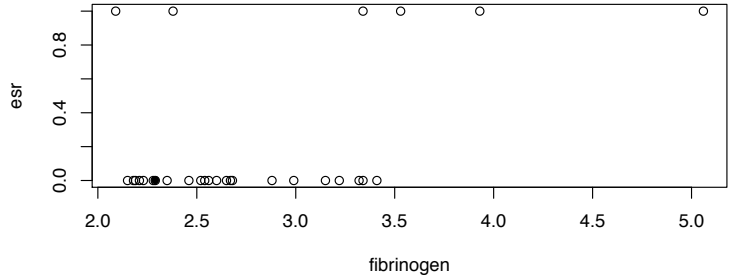
mesurés sur n individus.

Remarque : La loi de Y_i est entièrement déterminée par

$$f(x_i) = E(Y_i|x_i) = P(Y_i = 1|x_i).$$

M1 Isifar Modèles de régression Chapitre 1 21 / 24





Plan de ce cours

1. Introduction
2. Le modèle linéaire Gaussien, et quelques généralisations.
3. Le modèle linéaire généralisé, et en particulier la régression logistique.
4. La régression non paramétrique

Démarche de modélisation :

1. Regarder les données.
2. Choisir le type de modèle (paramétrique ou non, linéaire ou non...) puis préciser les hypothèses.
3. Ajuster le modèle.
4. Valider le modèle.
5. Selon les besoins, faire de l'inférence (tests, régions de confiance...), de la prédiction etc.

Le modèle retenu devra en outre respecter le principe de parcimonie.