Université Paris 7- Denis Diderot

Notes de cours pour

le Module

DATA MINING

Dominique Picard[1]

Chapitre 1

Introduction aux modèles de régression

[1]

Une première citation de H.G. Wells (1866-1946) : 'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.'

Une deuxième citation de Hal Varian, The McKinsey Quarterly, January 2009 : "I keep saying the sexy job in the next ten years will be statisticians. "

Le modèle de régression est probablement le modèle le plus vaste et le plus utilisé et étudié (encore maintenant des milliers d'articles paraissent dans des revues mathématiques chaque année sur le sujet) en statistique.

Il consiste à proposer une modélisation dans le cas de figure suivant. Pour employer un vocabulaire d'économiste, on dispose d'une variable endogène ou **expliquée** que l'on note généralement Y et d'un certain nombre p de variables exogènes ou **explicatives** que l'on note généralement $X^1, \ldots, X^p$. Les variables X^j pour $j = 1, \ldots, p$ apparaissant comme les causes d'un phénomème et la variable Y comme une conséquence, on a envie d'écrire qu'il existe une relation fonctionnelle entre la variable Y et les variables X^j pour $j = 1, \ldots, p$ soit

$$Y = f(X^1, \ldots, X^p)$$

pour une certaine fonction f sur laquelle on veut avoir des informations.

Le but de ce cours est d'étudier les principales méthodes d'estimation de cette fonction f lorsqu'on dispose de n données sur les variables $Y, X^1, \ldots, X^p$. Nous serons amenés à distinguer le cas où $p < n$ du cas où $p >> n$, plus difficile mais aussi très important dans le cadre actuel marqué plutot par la surabondance des données.

Suivant les hypothèses que l'on est pret à faire a priori, plusieurs méthodes seront envisagées. En particulier nous envisagerons plusieurs types de modèles.

- le modèle linéaire : f est supposée être une fonction linéaire. On s'intéresse à des variables X^j quantitatives.
- L'ANOVA ou le modèle logistique : f est supposée être linéaire. On s'intéresse à des variables X^j qualitatives ou catégorielles.
- le modèle de classification où les variables X^j sont quantitatives et où la variable Y est qualitative.
- le modèle paramétrique : f est supposée dépendre d'un paramètre θ inconnu. Mais la forme $f := f_\theta$ est connue.
- le modèle non paramétrique : f est supposée être complêtement inconnue. Cependant, on suppose qu'elle admet une certaine régularité.

Ces modèles sont très utilisés dans la pratique et dans de nombreux domaines. Donnons quelques exemples.

- Dans le domaine de l'économie : En vue d'une politique de relance par la consommation, on veut connaitre l'influence du revenu sur la consommation. Soit R le revenu d'un ménage et C sa consommation. L'INSEE modélise généralement la relation entre R et C par un modèle linéaire

$$R = a + bC.$$

  Le paramètre a représente la consommation incompressible d'un ménage (même sans revenu) et le paramètre b est appelé la **propension marginale à consommer**. Une estimation de b proposée par l'INSEE est environ 0.8.
- Dans le domaine de la biomédecine : On veut évaluer le risque d'apparition d'un cancer selon que la personne a été (ou non) exposée au tabac. La variable explicative est ici X qui prend 2 valeurs ("Fumeur" ou "NonFumeur") et la variable à expliquer est Y qui est une probabilité de risque (valeur comprise entre 0 et 1). On propose comme modèle

$$\text{Logit}(Y) = a + bX.$$

- Dans le domaine de l'environnement : il s'agit de prévoir la concentration d'ozone à partir des variables suivantes : force du vent, température et concentration d'oxyde d'azote. La forme particulière de la fonction f_θ est donnée par des physiciens qui utilisent des équations provenant de la mécanique des fluides.
- En signal : On enregistre un concert. On discrétise le signal en échantillonnant toutes les secondes. On note Y_i le signal recu au temps i. Ce signal est fonction du temps et on modélise par

$$Y_i = f(i) + \epsilon_i$$

  où ϵ_i contient tous les "bruits" enregistrés mais indésirables (les toux des gens, le bruit de la ventilation, ect..).

Chapitre 2

Modèle de régression linéaire

[1]

Ce modèle de régression est le plus utilisé et le mieux connu de toutes les personnes traitant des données dans des domaines divers.

2.1 Description du modèle

Soit Y la variable que l'on veut expliquer grâce aux p variables explicatives $X^1, \dots X^p$. On note X la matrice $n \times p$ qui contient les échantillons des variables X^j pour $j = 1, \dots p$:

$$X = \left(X_i^j\right)_{1 \le i \le n, 1 \le j \le p}.$$

La modélisation dite de régression linéaire multiple est la suivante

$$Y_i = \beta_1 \, X_i^1 + \ldots + \beta_p \, X_i^p + \epsilon_i, \quad 1 \le i \le n$$

ce qui est équivalent, en écriture matricielle à

$$\begin{array}{ccccccc} Y & = & X & \beta & + & \epsilon \\ (n,1) & & (n,p) & (p,1) & & (n,1) \end{array} \tag{2.1}$$

avec :

1. β est un paramètre de $\mathcal{R}^p$ inconnu et non aléatoire.
2. on impose au vecteur aléatoire ϵ de $\mathcal{R}^N$:
   - centrage : $E(\epsilon) = 0_n$.
   - indépendance et homoscédasticité : notons Σ la matrice de variance-covariance de ϵ. Alors $\Sigma = \sigma^2 Id_n$ pour $\sigma^2 > 0$ inconnu, déterministe.

   On ne connait pas forcément la loi de ϵ. On appelle ϵ l'erreur ou la perturbation.

Remarquons qu'en géneral, la constante 1_n de R^n fait partie des régresseurs (par défaut dans les logiciels). Le modèle est dit linéaire car il est linéaire en les paramètres β_j pour $j = 1, \dots, p$.

La plupart du temps dans ce cours, nous ferons l'hypothèse que les ε_i sont i.i.d. de loi normale $N(0, \sigma^2)$.

Une fois la modélisation choisie, il s'agit d'estimer les paramètres inconnus β, σ^2 du modèle (il y en a donc $p + 1$ au total).

1. Copyright @ 2013 Universite Paris-Diderot Dominique Picard

2.1.1 Exemples

1. Comparaison de 2 populations de même variance : On dispose de 2 échantillons $Z_1, \ldots, Z_m$ i.i.d. $N(\mu_1, \sigma^2)$ et $Z'_1, \ldots, Z'_m$ i.i.d. $N(\mu_2, \sigma^2)$. On les concatène pour former le vecteur

$$Y = (Z_1, \ldots, Z_n, Z'_1, \ldots, X'_m)^* = (Y_1, \ldots, Y_{m+n})^*$$

   Si on considère la matrice X de taille $n \times 2$, telle que

$$X_1^1 = \ldots = X_n^1 = 1,\ X_{n+1}^1 = \ldots = X_{n+m}^1 = 0$$

$$X_1^2 = \ldots = X_n^2 = 0,\ X_{n+1}^2 = \ldots = X_{n+m}^2 = 1$$

   et le vecteur $\beta = (\mu_1, \mu_2)^*$, il est facile de mettre notre modèle sous la forme (2.1).

2. Droite de régression. Supposons que l'on sache par des arguments théoriques (agronomiques, biologiques, économiques, physiques,...) que 2 quantités x (par exemple le temps) et y (par exemple la taille d'un animal) sont liées par une équation affine de la forme $y = ax + b$, dont on veut identifier les coefficients a et b. Une façon de procéder est de mesurer y_i pour différentes valeurs de x_i (appelée variable contrôlée) et de modèliser les erreurs par des $N(0, \sigma^2)$ indépendantes. On a alors la représentation (2.1), avec

$$X_1^1 = x_1, \ldots, X_n^1 = x_n,$$

$$X_1^2 = \ldots = X_n^2 = 1,$$

$$\beta = (a, b)^*$$

   Cet exemple peut se généraliser en remplaçant la relation affine par une relation de la forme :

$$y = \sum_{j=0}^{p} \beta_j f_j(x)$$

   Une régression polynomiale s'obtient par exemple en prenant

$$f_0 = 1,\ f_1(x) = x, \ldots, f_p(x) = x^p$$

3. On appelle **Analyse de la variance** (Anova) le cas où la matrice X est uniquement constituée de 1 et de 0.

   Donnons un exemple : Dans des conditions de culture de référence (0), une variété de blé a un rendement moyen de μ. On la soumet, dans des parcelles expérimentales à un traitement à 2 facteurs :

   1er facteur (par exemple, un engrais) auquel, outre le niveau 0 de référence, on donne 2 niveaux, notés 1 et 2 (par exemple, 2 doses différentes d'engrais).

   2eme facteur (par exemple, un niveau d'ensoleillement) auquel on donne soit le niveau de référence 0 soit le niveau 1.

Le modèle de base choisi est le suivant :

$$y = \mu + \alpha_i + \beta_j$$

Il est dit additif : Le effets des facteurs s'ajoutent simplement sans interférences. α_i représente l'effet du 1er facteur au niveau $i = 0, 1, 2$, β_j représente l'effet du 2eme facteur au niveau $j = 0, 1$. $\alpha_0 = \beta_0 = 0$. Le terme additif signifie que les effets des 2 facteurs s'ajoutent. Il est clair qu'on aurait pu aussi rajouter "une interaction" de la forme γ_{ij}, mais par souci de simplicité, nous ne l'avons pas fait ici.
Le but est d'obtenir des informations (estimation ou test) sur les α_i et les β_j. Pour cela, on réalise une expérimentation : On divise un champs en parcelles numérotées (6, dans l'exemple qui suit). Sur chaque parcelle, on applique les facteurs à un niveau prescrit. La description des niveux affectés aux parcelles s'appelle le plan de l'expérience. Ici, il est donné par le tableau suivant.

Parcelle	1	2	3	4	5	6
Facteur 1	0	1	2	0	1	0
Facteur 2	0	0	0	0	0	1

Si l'on suppose que l'on modèlise le rendement sur chaque parcelle par un effet de type (3) auquel s'ajoute une erreur $N(0, \sigma^2)$, et si l'on suppose les erreurs indépendantes, on obtient une équation du type $Y = X\beta + \varepsilon$, où Y est le vecteur des rendements, ε est le vecteur des erreurs, $\beta = (\mu, \alpha_1, \alpha_2, \beta_1)^*$ et X est la matrice suivante

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

2.2 Méthode des Moindres Carrés Ordinaires

2.3 Estimation de β

Nous allons utiliser ici la méthode dite des moindres carrés : Pour cela, on introduit la fonction,

$$\gamma(\beta, Y) = \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2$$

Cette fonction mesure la distance dans $\mathbb{R}^n$ entre le vecteur Y et sa prédiction par $X\beta$. Il est relativement naturel de choisir comme estimateur de β, un point $\hat{\beta}$ rendant cette quantité minimum.

$$\hat{\beta} = \text{Argmin}\{\gamma(\beta, Y); \beta \in \mathbb{R}^p\}$$

2.3.1 Interprétation géométrique

Si β parcourt $\mathbb{R}^p$, $X\beta$ parcourt l'espace vectoriel V engendré, dans $\mathbb{R}^n$, par les colonnes de la matrice X :

$$V = X(\mathbb{R}^p) \subset \mathbb{R}^n$$

Comme $\gamma(\beta, Y) = \|Y - X\beta\|^2$, nécessairement $X\hat{\beta}$, existe, est unique puisque c'est la projection sur V de Y, $X\hat{\beta} = \mathrm{Proj}_V(Y)$. On en déduit que $\hat{\beta}$ existe aussi toujours, mais n'est unique que si X est injectif.

Proposition 1 *Si $p \leq n$, la matrice X, de dimension $n \times p$ est injective si et seulement si X^*X est inversible.*

Démonstration de la Proposition.

Il sufit de démontrer que $\ker(X) = \ker(X^*X)$. Il est clair que $\ker(X) \subset \ker(X^*X)$. Maintenant, soit $u \in \ker(X^*X)$, on a $X^*Xu = 0$, d'où $u^*X^*Xu = 0$, i.e. $\|Xu\|^2 = 0 \Longrightarrow Xu = 0 \Longrightarrow u \in \ker X$.

Résolution algébrique

$$\begin{array}{rcll} X\hat{\beta} = \mathrm{Proj}_V(Y) & \Longleftrightarrow & \langle Y - X\hat{\beta}, Xb\rangle = 0, & \forall b \in \mathbb{R}^p \\ & \Longleftrightarrow & b^*X^*Y = b^*X^*X\hat{\beta}, & \forall b \in \mathbb{R}^p \\ & \Longleftrightarrow & X^*Y = X^*X\hat{\beta} & \end{array}$$

D'oú, en utilisant la proposition si X est injective,

$$\hat{\beta} = (X^*X)^{-1}X^*Y$$

Remarque : Si X^*X n'est pas inversible, on n'a pas unicité de $\hat{\beta}$, mais existence. Donnons une solution, utilisant la pseudoinverse : X^*X étant une matrice symétrique, positive, elle s'écrit M^*DM avec M matrice orthogonale et D est une matrice diagonale, dont les coefficients diagonaux sont notés r_i^2. On suppose $r_i^2 > 0, \forall i = 1, \ldots, k,\ r_i^2 = 0, \forall i \geq k+1$. Appelons pseudoinverse de X^*X la matrice

$$(X^*X)^{(-1*)} = M^* \begin{pmatrix} \frac{1}{r_1^2} & \ldots & \ldots & 0 & 0 & 0 \\ & & \ldots & & & \\ 0 & \ldots & \frac{1}{r_k^2} & \ldots & 0 & 0 \\ & & \vdots & & & \\ 0 & \ldots & 0 & \ldots & 0 & 0 \end{pmatrix} M$$

Notons que si X^*X est inversible, alors pseudoinverse et inverse coincident. On vérifie facilement que

$$\hat{\beta} = (X^*X)^{(-1*)}X^*Y$$

est une solution de notre problème, et que l'opérateur de projection sur V est donné par :

$$X\hat{\beta} = X(X^*X)^{(-1*)}X^*Y = \mathrm{Proj}_V(Y)$$

$\triangle$

Rappelons que si $V^\perp$ est le supplémentaire orthogonal de V,

$$\mathrm{Proj}_{V^\perp}(Y) = Y - \mathrm{Proj}_V(Y) = [I_n - \mathrm{Proj}_V](Y) = [I_n - X(X^*X)^{-1}X^*]Y$$

Définition 1 *On appelle vecteur des résidus, le vecteur*

$$\hat{\varepsilon} = [I_n - X(X^*X)^{-1}X^*]Y.$$

Il représente l'erreur de prédiction. Le carré de sa norme s'appelle l'erreur quadratique.

Exemples :

1. Dans le cas élémentaire suivant :

$$Y_i = \mu + \varepsilon_i$$

l'estimateur des moindres carrés se calcule facilement et vaut $\bar{Y}_n = \frac{\sum_{i=1}^n Y_i}{n}$.

2. Dans le cas d'une régression linéaire, nous avons vu que $\beta = (a, b)^*$ et

$$X = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

De sorte que

$$X^*X = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix}$$

Dans ce cas, un changement de paramètres peut rendre les choses plus aisées : En effet, si on introduit $\bar{x}_n = \frac{\sum_{i=1}^n x_i}{n}$, le modèle s'ecrit :

$$Y_i = az_i + b' + \varepsilon_i,\ z_i = x_i - \bar{x}_n,\ b' = b + \bar{x}_n$$

et clairement minimiser $\sum_{i=1}^n (Y_i - az_i + b')^2$ équivaut à minimiser $\sum_{i=1}^n (Y_i - ax_i + b)^2$, avec la relation suivante $\hat{b}' = \hat{b} + \hat{a}\bar{x}_n$. L'équation (2) introduit un nouveau modèle linéaire dont la matrice X' s'écrit :

$$X'^*X' = \begin{pmatrix} \sum_{i=1}^n z_i^2 & 0 \\ 0 & n \end{pmatrix}$$

Cette matrice est inversible si et seulement si $\sum_{i=1}^n z_i^2 \neq 0$, c'est à dire si les x_i ne sont pas tous égaux. Dans ce cas, on obtient facilement :

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)Y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2},\ \hat{b} = \bar{Y}_n + \hat{a}\bar{x}_n$$

3. Considérons maintenant la régression périodique suivante :

$$Y_i = a_0 + a_1 \cos(2\pi \frac{i}{n}) + a_2 \sin(2\pi \frac{i}{n}) + \varepsilon_i,\ i = 1, \ldots, n$$

On vérifie que en utilisant les relations sur les racines de l'unité que X^*X se met sous la forme suivante :

$$\begin{pmatrix} n & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) & \sum_{i=1}^n \sin(2\pi \frac{i}{n}) \\ \sum_{i=1}^n \cos(2\pi \frac{i}{n}) & \sum_{i=1}^n \cos(2\pi \frac{i}{n})^2 & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) \sin(2\pi \frac{i}{n}) \\ \sum_{i=1}^n \sin(2\pi \frac{i}{n}) & \sum_{i=1}^n \cos(2\pi \frac{i}{n}) \sin(2\pi \frac{i}{n}) & \sum_{i=1}^n \sin(2\pi \frac{i}{n})^2 \end{pmatrix} = \begin{pmatrix} n & 0 & 0 \\ 0 & \frac{n}{2} & 0 \\ 0 & 0 & \frac{n}{2} \end{pmatrix}$$

On en déduit que

$$\hat{a}_0 = \bar{Y}_n,\ \hat{a}_1 = \sum_{i=1}^n \cos(2\pi \frac{i}{n}) Y_i,\ \hat{a}_2 = \sum_{i=1}^n \sin(2\pi \frac{i}{n}) Y_i$$

△

2.3.2 Calcul récursif, Méthode de Gram Schmidt

Nous proposons ici une méthode pour calculer $\hat{\beta}$ de façon récursive. Appelons X^j la colonne numéro j de la matrice X pour $1 \leq j \leq p$.

Considérons le cas suivant dans lequel les MCO sont particulièrement faciles à calculer : Supposons que les **colonnes** de X soient orthogonales (i.e. X^tX est une matrice diagonale dont les coeficients diagonaux sont les carrés des normes des colonnes : $\sum_{i=1}^n [X_i^j]^2 = \langle X^j, X^j \rangle$. Dans ce cas, les coefficients $\hat{\beta}_j$ valent simplement :

$$\hat{\beta}_j = \frac{\langle X^j, Y \rangle}{\langle X^j, X^j \rangle}$$

Rappelons nous maintenant le procédé d'orthonormalisation de Gram Schmidt qui pour des vecteurs quelconques $u_1, \ldots, u_k$ (tels que l'espace engendré par ces vecteurs ($sp\ \{u_1, \ldots, u_k\}$) soit de dimension k) introduit les vecteurs $v_1, \ldots, v_k$ qui sont orthogonaux et vérifient $sp\ \{u_1, \ldots, u_l\} = sp\ \{v_1, \ldots, v_l\}$, pour tout $1 \leq l \leq k$. Ce procédé consiste simplement à construire les v_l sous la forme suivante : $v_1 = u_1$,

$$v_\ell = u_\ell - P_{v_{\ell-1}} u_\ell - \ldots - P_{v_1} u_\ell, \quad \ell \geq 2.$$

(P_{v_j} désigne la projection sur le vecteur v_j).

Remarquons que pour $1 \leq j \leq \ell - 1$,

$$P_{v_j} u_\ell = \frac{\langle v_j, u_\ell \rangle}{\langle v_j, v_j \rangle}.$$

De plus comme les v_j sont orthogonaux, $P_{v_{\ell-1}} u_\ell + \ldots + P_{v_1} u_\ell$ est la projection de u_ℓ sur l'espace $sp\{v_1, \ldots, v_{\ell-1}\}$. Donc v_l est en fait le 'résidu' de la projection de la projection de u_ℓ sur l'espace $sp\{v_1, \ldots, v_{\ell-1}\}$.

Considérons maintenant, dans le cas $p \leq n$ et où la matrice X est de rang p, l'algorithme suivant :

– Initialisation : $Z^1 = X^1$
– Pour $l = 2$ jusqu'à p calculer : Z^l le résidu de la projection de X^l sur $Z^{l-1}, \ldots, Z^1$, i.e.

$$Z^l = X^l - \frac{\langle Z^{l-1}, X^l \rangle}{\langle Z^{l-1}, Z^{l-1} \rangle} Z^{l-1} - \ldots - \frac{\langle Z^1, X^l \rangle}{\langle Z^1, Z^1 \rangle} Z^1.$$

Montrer qu'alors

$$\hat{\beta}_p = \frac{\langle Z^p, Y \rangle}{\langle Z^p, Z^p \rangle}.$$

En changeant l'ordre des colonnes de la matrice X, on peut s'arranger pour faire apparaitre X^j en dernier pour chaque j. Cela donne une façon de calculer les $\hat{\beta}_j$ sans inverser la matrice. (Attention on a donc p calculs différents.)

Cet algorithme permet aussi de mesurer les problèmes qui peuvent arriver au cours d'une telle estimation. Supposons en effet que le vecteur X^p soit très corrélé avec (par exemple) X^{p-1} (ou soit proche d'une combinaison linéaire de $X^1, \ldots, X^{p-1}$) ; dans ce cas le résidu Z_p va être très petit et par voie de conséquence l'estimation de $\hat{\beta}_p$ très instable.

2.4 Lois des estimateurs. Estimation de σ^2.

Nous allons maintenant montrer la proposition suivante sous l'hypothèse que les ε_i sont i.i.d. $N(0, \sigma^2)$:

Proposition 2 *Sous la condition, $p \leq n$, X^*X inversible, le vecteur de dimension $p + n$:*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix}$$

est un vecteur gaussien de moyenne et variance :

$$\begin{pmatrix} \beta \\ 0 \end{pmatrix}, \quad \sigma^2 \begin{pmatrix} (X^*X)^{-1} & 0 \\ 0 & I_n - X(X^*X)^{-1}X^* \end{pmatrix}$$

Preuve de la Proposition

Espérances et variances de $\hat{\beta}$ Dans ce paragraphe, l'hypothèse de gaussiannité sur les ε_i est inutile. Les résultats sont encore vrais si l'on suppose que $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}\varepsilon = \sigma^2 I_n$.

Comme $\hat{\beta} = (X^*X)^{-1}X^*Y$, on a $\mathbb{E}\hat{\beta} = \mathbb{E}(X^*X)^{-1}X^*(X\beta + \varepsilon) = \beta$.

D'autre part,

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= (X^*X)^{-1}X^*[\mathrm{Var}(Y)]X(X^*X)^{-1} \\ &= (X^*X)^{-1}X^*[\mathrm{Var}(\varepsilon X)](X^*X)^{-1} \\ &= \sigma^2(X^*X)^{-1}X^*X(X^*X)^{-1} = \sigma^2(X^*X)^{-1}. \end{aligned}$$