Modèle linéaire en statistique et extensions

C. Ambroise

Laboratoire de Mathématiques et Modélisation d'Évry
UMR CNRS 8071

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

Régression linéaire multiple (x_j quantitatif), Analyse de la variance (ANOVA) (x_j qualitatifs) sont les deux exemples les plus courants

- y est une variable aléatoire normale,
- les bruits ϵ sont indépendants,
- linéarité en les paramètres

On veut comparer 3 types d'engrais. Pour ceci, on tire au hasard trois ensembles de parcelles sur lesquelles on utilise chacun des engrais. On mesure le rendement obtenu sur chacune des parcelles :

- échantillon 1 (engrais 1) : 35, 34, 33, 36, 37
- échantillon 2 (engrais 2) : 41, 38, 40, 40, 41
- échantillon 3 (engrais 3) : 35, 37, 34, 39, 35.

Le choix de l'engrais a-t-il une influence sur le rendement ?

$$
\begin{aligned}
\mathit{rendement} = & \mathit{effet.global} + && (1)\\
& \mathit{effet.engrais}_1\ x_1 + && (2)\\
& \mathit{effet.engrais}_2\ x_2 + && (3)\\
& \mathit{effet.engrais}_3\ x_3 + \mathit{bruit} && (4)
\end{aligned}
$$

Où les x_j sont des variables binaires.

1. modèle linéaire généralisé : Y non normale (loi de la famille exponentielle)
2. modèles à effets mixtes : bruit structuré
3. modèles non linéaires
4. modèles avec *a priori* sur les paramètres

Plan du cours I

4 cours et 2 séances de TD sur machines

lundi 09/02 Cours Rappel et régression linéaire multiple

lundi 16/02 Cours Modèle linéaire généralisé

lundi 02/03 TD Modèle linéaire

lundi 09/03 Cours Modèle mixte

lundi 16/03 Cours Modèles additifs et régularisés

lundi 24/03 TD Modèle linéaire généralisé et modèle additif (et contrôle en seconde partie)

Évaluation

1. Devoir maison (1/4 de la note)
2. Partiel (3/4 de la note)

Plan

Gaussienne univariée

La variable y est dite gaussienne de moyenne μ et variance σ^2 lorsqu'elle a pour densité de probabilité

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}. \tag{5}$$

On note $\mathcal{N}(\mu, \sigma^2)$.

Rappel :

- Toute combinaison linéaire de variables gaussienne est gaussienne.

- Théorème de la limite centrale.

$\boldsymbol{y}' = [y_1 y_2 \cdots y_p]$ est un vecteur normal de vecteur moyenne $\boldsymbol{\mu}$ et de matrice de variance covariance V lorsque le vecteur a pour densité de probabilité

$$f(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^p |V|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})' V^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}. \tag{6}$$

Si l'on considère p variables gaussiennes indépendantes et identiquement distribuées, la loi jointe des p variables devient :

$$f(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^p \sigma^p} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})' V^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}. \tag{7}$$

avec V^{-1} une matrice diagonale avec des $1/\sigma^2$ sur la diagonale.

Si K est une $p \times p$ matrice de changement de base (inversible) et $\boldsymbol{x}$ un vecteur gaussien dont toutes les composantes sont indépendantes et identiquement distribuées de variance unité. Quelle est la distribution de $\boldsymbol{y} = K\boldsymbol{x}$? Il suffit de faire le changement de variables pour obtenir :

$$g(y_1, y_2, ..., y_p) = f(K^{-1}\boldsymbol{y} = (x_1, ..., x_p)^t) J(y_1, ..., y_p)$$

où $J(y_1, ..., y_p)$ est le jacobien de la transformation $K^{-1}\boldsymbol{y} = (x_1(\boldsymbol{y}), ..., x_p(\boldsymbol{y}))^t$:

$$J(y_1, ..., y_p) = mod \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 & ... & \partial x_1/\partial y_p \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 & ... & \partial x_2/\partial y_p \\ \vdots & \vdots & & \vdots \\ \partial x_p/\partial y_1 & \partial x_p/\partial y_2 & ... & \partial x_p/\partial y_p \end{vmatrix}.$$

Dans ce cas on a :

- $$E(\boldsymbol{y}) = E(K\boldsymbol{x}) = K\boldsymbol{\mu}_x = \boldsymbol{\mu}$$
- $$var(\boldsymbol{y}) = var(K\boldsymbol{x}) = Kvar(\boldsymbol{x})K^t = KK^t$$
- $$(\boldsymbol{x}-\mu)^t(\boldsymbol{x}-\mu) = (\boldsymbol{y}-\boldsymbol{\mu}_y)^t(K^{-1})^tK^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y) = (\boldsymbol{y}-\boldsymbol{\mu}_y)^t var(\boldsymbol{y})^{-1}($$
- $$J(y_1, ..., y_p) = |det(K^{-1})| = |det(var(\boldsymbol{y}))|^{1/2} = |det(V)|^{-1/2}$$
- $\boldsymbol{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, KK^t) : f(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^p |V|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})' V^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}$.

Soit $\boldsymbol{x}$ un vecteur gaussien de p-vecteur moyenne nul et matrice de variance identité : $\boldsymbol{x} \sim \mathcal{N}_p(0, I_p)$.
Soit $\boldsymbol{y}$ les coordonnées de $\boldsymbol{x}$ dans une base orthonormée.
On a $\boldsymbol{y} = A\boldsymbol{x}$ où A est la matrice dont les colonnes sont les coordonnées des vecteurs canoniques dans la nouvelle base.
Dans ce cas $\boldsymbol{y} \sim \mathcal{N}_p(0, AA^t = I_p)$.

Plan

Soit $x_1, ..., x_k$ k variables gaussiennes centrées réduites :

$$Q = \sum_{i=1}^{k} x_i^2,$$

est distribuée selon une loi du Chi 2 à k degré de liberté. On note :

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi^2_k.$$

Si $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ alors $\boldsymbol{y} = \Sigma^{-1/2}(\boldsymbol{x} - \mu) \sim \mathcal{N}_p(0, I)$ et

$$Q = \boldsymbol{y}^t \boldsymbol{y} \sim \chi^2_p$$

.
L'équation

$$P(Q \leq q) = \alpha$$

avec $q = \chi^2_{p,\alpha}$ définit une éllipsoïde d'équiprobabilité

Plan

Theorem

Soit $\boldsymbol{x}$ un vecteur gaussien de p-vecteur moyenne nul et matrice de variance identité : $\boldsymbol{x} \sim \mathcal{N}_p(0, I_p)$.
Soit E un espace vectoriel de dimension $r < p$ et P_E la projection orthogonale sur E et $P_{E^\perp}$ la projection sur le complémentaire de E.
Les projections $P_E(\boldsymbol{x})$ et $P_{E^\perp}(\boldsymbol{x})$ sont des vecteurs gaussiens indépendants de lois

$$P_E \boldsymbol{x} \sim \mathcal{N}_p(0, P_E) \text{ et } P_{E^\perp} \boldsymbol{x} \sim \mathcal{N}_p(0, P_{E^\perp}).$$

Les normes des projections sont des variables aléatoires qui suivent des lois du χ^2 :

$$\|P_E \boldsymbol{x}\|^2 \sim \chi^2_r \text{ et } P_{E^\perp} \boldsymbol{x} \sim \chi^2_{p-r}.$$

Démonstration

Soit $B = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_p]$ une base orthonormée de $\mathcal{R}^p$ telle que $[\boldsymbol{b}_1 \cdots \boldsymbol{b}_r]$ soit une base orthonormée de E et $[\boldsymbol{b}_{r+1} \cdots \boldsymbol{b}_p]$ une base de $E^\perp$.

Notons P la matrice de passage de la base canonique à la base B.

Soit $\boldsymbol{y} = y_1, ..., y_r$ les coordonnées de $\boldsymbol{x}$ dans la base orthonormée $[\boldsymbol{b}_1 \cdots \boldsymbol{b}_r]$ de E :

$$P_E(\boldsymbol{x}) = y_1\boldsymbol{b}_1 + ... + y_r\boldsymbol{b}_r = J_r\boldsymbol{y}, \tag{8}$$
$$P_{E^\perp}(\boldsymbol{x}) = y_{r+1}\boldsymbol{b}_{r+1} + ... + y_p\boldsymbol{b}_p = J_{p-r}\boldsymbol{y}, \tag{9}$$

où J_r est une matrice diagonale avec des 1 sur les r premiers coefficients diagonaux et des 0 ensuite, et $J_{p-r} = I_n - J_r$.
On a par définition $\boldsymbol{y} = P^t\boldsymbol{x}$. Et d'après la remarque de la sous-section vecteur gaussien et base orthonormée : $\boldsymbol{y} \sim \mathcal{N}_p(0, I_p)$. Les deux projections sont donc indépendantes car combinaisons linéaires de variables gaussiennes indépendantes.
On peut alors revenir au vecteur $\boldsymbol{x}$ en remarquant que $P_E\boldsymbol{x} = PJ_r\boldsymbol{y}$ et $P_{E^\perp}\boldsymbol{x} = PJ_{p-r}\boldsymbol{y}$ gaussiens centrés indépendants de matrice de covariance respectives P_E et $P_{E^\perp}$.

Et si l'on considère la norme de la projection :

$$\|P_E(\boldsymbol{x})\|^2 = \sum_{j=1}^{r} y_j^2 \sim \chi_r^2 \tag{10}$$

$$\|P_{E^\perp}(\boldsymbol{x})\|^2 = \sum_{j=r+1}^{p} y_j^2 \sim \chi_{(p-r)}^2 \tag{11}$$

Soit $\boldsymbol{x} = (x_1, ..., x_n)$, n variables gaussiennes indépendantes de loi $\mathcal{N}(0, \sigma^2)$.
On a $\boldsymbol{x} - \mu\mathbb{I}_n \sim \mathcal{N}(0, \sigma^2 I_n)$.
Soit E l'espace vectoriel engendré par le vecteur $\frac{1}{\sqrt{n}}\mathbb{I}_n = \frac{1}{\sqrt{n}}(1, ..., 1)^t$ et $E^\perp$ son complémentaire dans $\mathbb{R}^n$.
La projection de $\boldsymbol{x} - \mu\mathbb{I}_n$ sur E est le vecteur :

$$P_E(\boldsymbol{x} - \mu\mathbb{I}_n) = \frac{1}{\sqrt{n}}\mathbb{I}_n^t(\boldsymbol{x} - \mu)\frac{1}{\sqrt{n}}\mathbb{I}_n = (\bar{x} - \mu)\mathbb{I}_n$$

La projection de $\boldsymbol{x} - \mu\mathbb{I}_n$ sur $E^\perp$ est le vecteur :

$$P_{E^\perp}(\boldsymbol{x} - \mu\mathbb{I}_n) = (\boldsymbol{x} - \mu\mathbb{I}_n) - P_E(\boldsymbol{x} - \mu\mathbb{I}_n) = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

Les deux projections sont indépendantes (orthogonales). Les normes des projections sont des variables qui suivent des $\sigma^2\chi^2$ de degré respectifs 1 et $n-1$:

$$n(\overline{x}-\mu)^2 \sim \sigma^2\chi^2_1,$$

$$\sum (x_i-\overline{x})^2 \sim \sigma^2\chi^2_{n-1}.$$

Plan

La régression est utile pour expliquer ou modéliser la relation entre une variable y, appelée :

- variable à expliquer,
- réponse,
- sortie,

et des variables $x_1, x_2, ...x_p$ appelées

- variables explicatives,
- variables d'entrées.

La régression peut serivr différent buts

1. prédire des observations futures

2. confirmer une relation entre variables explicatives et variables à expliquer,

3. décrire une structure de données (analyse exploratoire).

- Au 19ème siècle, navigation utilisant la position des étoiles : Legendre méthodes des moindres carrés

- contestation de la paternité par Gauss, qui affirme être l'inventeur de la méthode et montre son optimalité pour des erreurs Gaussiennes.

- méthode baptiséee par Sir Francis Galton (1870), le cousin de Darwin, à cause d'une étude sur la régression de la taille des fils par rapport à leur père

Plan

Supposons que l'on ait observé les variables $x_1, \ldots, x_p$, Y pour n individus (dans n situations différentes). Les données se présentent donc sous la forme suivante :

$$\begin{array}{cccc} x_{11} & \ldots & x_{1p} & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} & y_n \end{array}$$

On suppose que chaque valeur observée y_i sur un individu i est une réalisation d'une v.a.r. Y_i de la forme :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad i = 1, n$$

avec $\mathbb{E}(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma^2$ et $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

Remarques

- dans le modèle linéaire ce sont les paramètres qui définissent la linéarité.

- Les prédicteurs ne sont pas forcement linéaires.

Matriciellement, ces n équations s'écrivent

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon$$

avec

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta \\ \vdots \\ \beta \end{pmatrix}$$

et

$$\mathbb{E}(\varepsilon) = 0 \quad \text{var}(\varepsilon) = \sigma^2 I_n.$$

Ce modèle étant posé, nous allons successivement aborder les problèmes suivants :

- estimation des paramètres $\boldsymbol{\beta}$ et σ^2 ;
- tests d'hypothèses relatives aux paramètres (« significativité »de la régression, etc.) ;
- prédiction de Y ou $\mathbb{E}(Y)$ pour une nouvelle valeur de $\boldsymbol{x}$;
- diagnostic de la régression (validation du modèle) ;
- sélection d'un ensemble de variables explicatives « pertinentes ».

Soit $\hat{\beta}$ un estimateur du paramètre vectoriel β. La méthode des moindres carrés consiste à choisir $\hat{\beta}$ de façon à minimiser la somme des carrés des écarts entre les observations Y_i et les prédictions $\widehat{Y}_i = \boldsymbol{x}_i'\hat{\beta}$. On cherche donc à minimiser le critère

$$\begin{aligned} RSS(\hat{\beta}) &= \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i'\hat{\beta})^2 \\ &= (\mathbf{Y} - X\hat{\beta})'(\mathbf{Y} - X\hat{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'X'\mathbf{Y} + \hat{\beta}'X'X\hat{\beta}. \end{aligned} \tag{12}$$

Theorem

Le minimum de la fonction $RSS(\hat{\beta})$ *est obtenu pour*

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{Y},$$

c'est-à-dire que l'on a

$$RSS(\hat{\beta}) = \min_{\beta} RSS(\beta).$$

$\hat{\beta}$ *est appelé* estimateur des moindres carrés *de* β*.*

Preuve : Il s'agit d'une fonction de $p+1$ variables. Pour en trouver le minimum, il suffit d'annuler le gradient, c'est-à-dire le vecteur des dérivées partielles :

$$\nabla S = \left(\frac{\partial RSS}{\partial \beta_0}, \ldots, \frac{\partial S}{\partial \beta_p} \right),$$

ce qui conduit à un système de $p+1$ équations à $p+1$ inconnues. Il vient ici

$$\nabla RSS = -2X'\mathbf{Y} + 2X'X\beta = 0. \tag{13}$$

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

On a

$$\hat{\mathbf{Y}} = X\hat{\beta} = X(X'X)^{-1}X'\mathbf{Y} = H\mathbf{Y}$$

en notant $H = X(X'X)^{-1}X'$. Cette matrice H a des propriétés remarquables. En effet, H est symétrique (évident), et de plus

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H.$$

La matrice H est donc idempotente (c'est un opérateur de projection orthogonale, comme nous le verrons par la suite). De même, on peut écrire

$$\widehat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (I_n - H)\mathbf{Y} = R\mathbf{Y}$$

avec $R = I_n - H$. On vérifie aisément que R a les mêmes propriétés que H (symétrie et idempotence) : c'est également un opérateur de projection orthogonale.

Plaçons nous dans $\mathbb{R}^n$ (cf. figure 1) et considérons les vecteurs

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \; j = 1, p \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Le modèle linéaire s'écrit avec ces notations

$$\mathbf{Y} = \beta_0 \mathbf{1} + \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j + \varepsilon.$$

La méthode des moindres carrés peut être interprétée comme la recherche de la meilleure approximation de $\mathbf{Y}$ dans le sous-espace $\mathcal{L}$ de $\mathbb{R}^n$ engendré par les $p+1$ vecteurs $\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. On cherche en effet

$$\hat{\mathbf{Y}} = \widehat{\beta}_0 \mathbf{1} + \sum_{j=1}^{p} \widehat{\beta}_j \boldsymbol{x}_j \in \mathcal{L}$$

tel que la distance euclidienne $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ soit minimum. On sait que la solution consiste à définir $\hat{\mathbf{Y}}$ comme la projection orthogonale de $\mathbf{Y}$ sur $\mathcal{L}$. On a vu en effet que

$$\hat{\mathbf{Y}} = H\mathbf{Y},$$

H étant un opérateur de projection orthogonale.

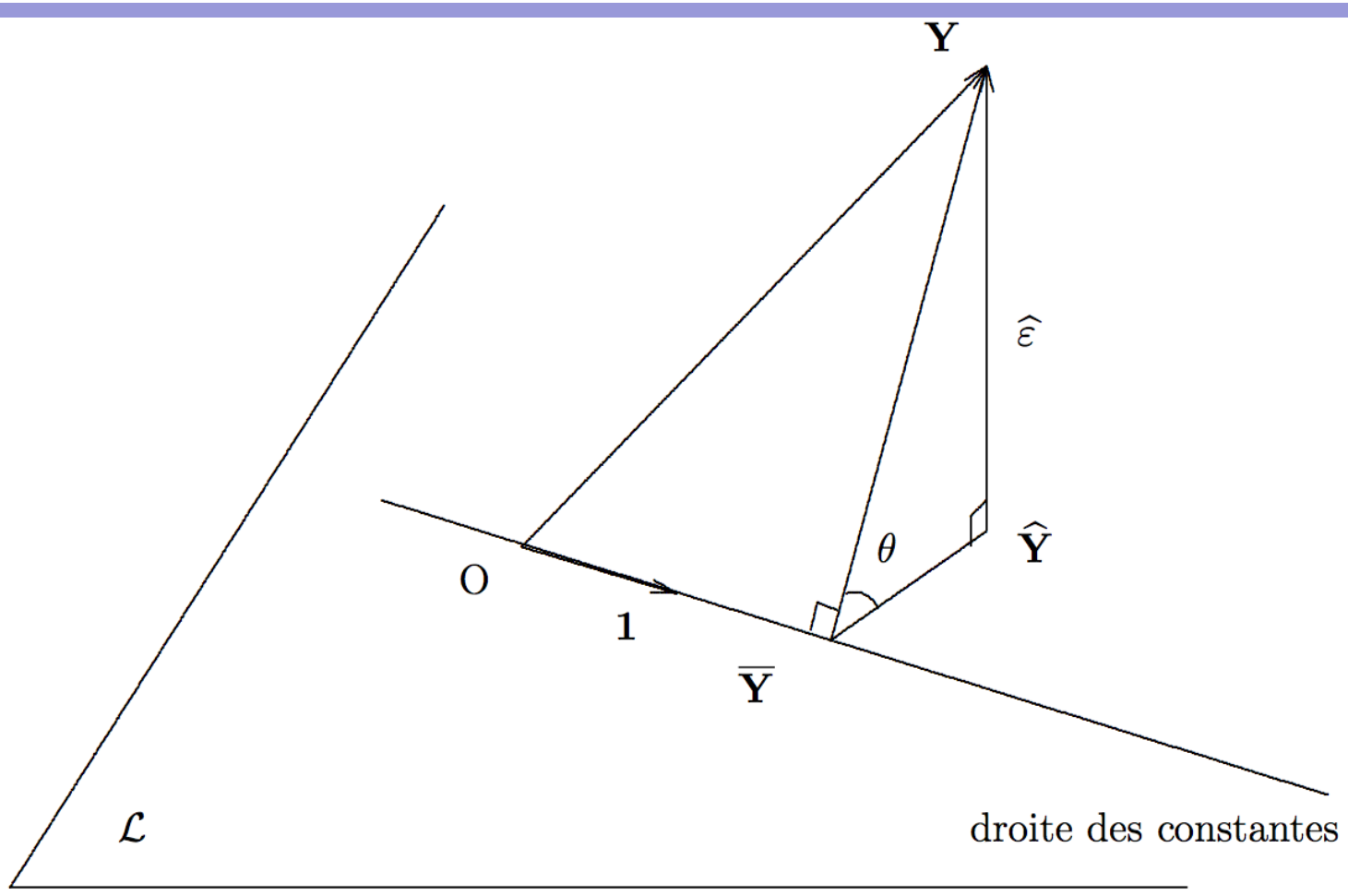

FIGURE : Interprétation géométrique de la régression linéaire

Cette représentation géométrique permet de retrouver sans calculs fastidieux plusieurs résultats intéressants. Tout d'abord, on a

$$\widehat{\varepsilon} \perp \mathbf{1} \Rightarrow \sum_{i=1}^{n} \widehat{\varepsilon}_i = 0,$$

d'où l'on déduit

$$\frac{1}{n} \sum_i \widehat{Y}_i = \frac{1}{n} \sum Y_i = \overline{Y}.$$

Par ailleurs, la projection orthogonale de $\mathbf{Y}$ sur l'axe dirigé par $\mathbf{1}$ a pour coordonnée

$$\frac{< \mathbf{Y}, \mathbf{1} >}{\|\mathbf{1}\|^2} = \overline{Y}.$$

Il en est de même, d'après ce qui précède, pour la projection orthogonale de $\widehat{\mathbf{Y}}$ sur $\mathbf{1}$. Enfin, on a de manière évidente :

$$\widehat{\mathbf{Y}} \perp \widehat{\varepsilon}.$$

Notons $\overline{\mathbf{Y}} = \overline{Y}\mathbf{1}$. En appliquant le théorème de Pythagore au triangle $(\mathbf{Y}, \hat{\mathbf{Y}}, \overline{\mathbf{Y}})$, on obtient finalement la relation très importante suivante, appelée *équation d'analyse de la variance* :

$$\|\mathbf{Y} - \overline{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 + \|\widehat{\varepsilon}\|^2,$$

ce que l'on peut encore écrire, en divisant par n :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

soit encore

$$S_{YY} = S_{reg} + RSS.$$

Cette équation est appelée *équation d'analyse de la variance*. Le terme de gauche (S_{YY}) est la variance empirique des Y_i, il caractérise la dispersion des valeurs observées de la variable à expliquer. Le premier terme du membre de droite (S_{reg}) est la variance empirique des $\widehat{Y}_i$, que l'on appelle variance expliquée par la régression. Le second terme du membre de droite (RSS) est la variance des résidus, ou variance résiduelle.

- S_{YY} dépend de n quantités $Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}$ liées par la relation

$$\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0.$$

  Ce terme a donc $n - 1$ d.d.l.

- On a $\widehat{Y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ et $\overline{Y} = \overline{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}$. Par conséquent, le terme S_{reg} est fonction des paramètres $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ (le terme $\widehat{\beta}_0$ s'annule dans chacune des différences $\widehat{Y}_i - \overline{Y}$). La variance expliquée a donc p d.d.l.

- Par conséquent, le nombre de d.d.l associé à la variance résiduelle est $n - p - 1$.

La plupart des logiciels statistiques présentent les résultats de la régression sous forme d'un tableau (appelé *tableau d'analyse de la variance*), où figurent les différents termes de l'équation d'analyse de la variance, et les nombres de d.d.l associés (cf. tableau 6).

source de variation	d.d.l.	SS	MS=SS/d.d.l.
régression	p	$\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	$\frac{1}{p}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$
résiduelle	$n-p-1$	$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$\frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \widehat{\sigma}^2$
totale	$n-1$	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

TABLE : Tableau d'analyse de la variance (SS : *sum of squares* ; MS : *mean square*).

$\hat{\boldsymbol{\beta}}$ est un estimateur BLUE (Best Linear Unbiased EStimate). C'est l'estimateur de plus petite variance parmi tous les estimateurs sans biais $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

Espérance et variance

Plan

Si l'on ajoute l'hypothèse que les résidus sont Gaussiens :

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

on a aussi d'après la section précédente

$$\hat{\beta} \sim \mathcal{N}_n(\beta, \sigma^2 (X'X)^{-1})$$

Comme H est une matrice de projection dans un espace à $p+1$ dimension et que

$$\hat{\mathbf{Y}} \perp \widehat{\varepsilon}$$

d'après le théorème de Cochran on a

$$\|\widehat{\varepsilon}\|^2 \sim \sigma^2 \chi^2_{n-p-1}$$

On peut donc estimer σ^2 par

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\widehat{\varepsilon}\|^2 = \frac{1}{n-p-1}\sum_i (Y_i - \hat{Y}_i)^2$$

Comparaison de modèles

Une question légitime lorsque l'on considère un modèle est

> *Toutes les variables explicatives sont nécessaires pour expliquer la variable de sortie ?*

Une façon de poser la question consiste à considérer le test d'hypothèse suivant

$$\begin{cases} H_0 : \text{un petit modèle } \omega \text{ est adapté} \\ H_1 : \text{un grand modèle } \Omega \text{ est adapté} \end{cases}$$

avec $\omega \subset \Omega$

Aperçu géométrique de la comparaison de modèles

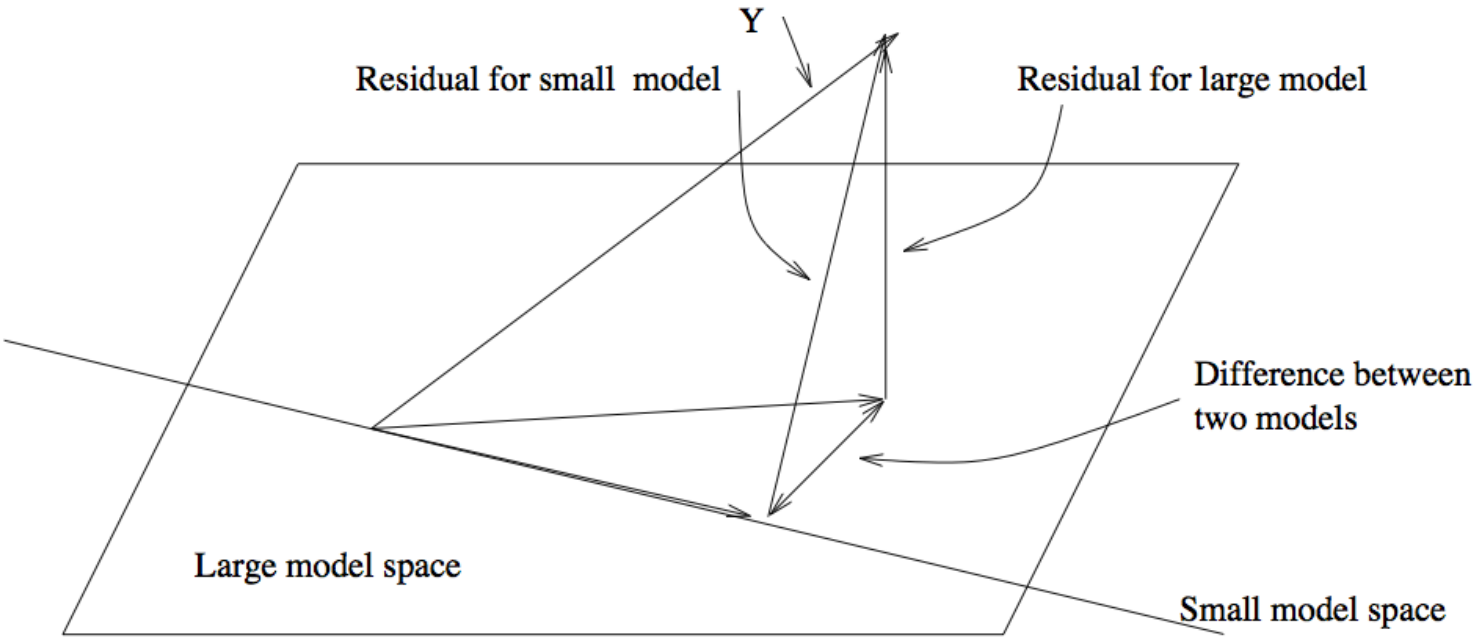

FIGURE : Figure du livre de Faraway : 'Pratical regression and anova using R'

Principe de comparaison

On choisira H_1 (le modèle le plus grand Ω) si

- les résidus du grand modèle sont vraiment petits comparés au modèle concurrent
- $$RSS_\Omega < RSS_\omega$$
- $$\frac{RSS_\omega - RSS_\Omega}{RSS_\Omega} >> 1$$
- $$F = \frac{RSS_\omega - RSS_\Omega}{RSS_\Omega} \frac{(n - q_\Omega)}{(q_\Omega - q_\omega)} > f$$

Statistique de décision

D'après le théorème de Cochran nous avons

$$\frac{RSS_\Omega}{\sigma^2} \sim \chi^2_{n-q_\Omega}$$

et comme les résidus des deux modèles emboités sont orthogonaux, il sont indépendants et Cochran nous donne

$$\frac{RSS_\omega - RSS_\Omega}{\sigma^2} \sim \chi^2_{q_\Omega - q_\omega}$$

Donc sous H_0

$$F = \frac{RSS_\omega - RSS_\Omega}{RSS_\Omega} \frac{(n - q_\Omega)}{(q_\Omega - q_\omega)} \sim \mathcal{F}_{(n-q_\Omega);(q_\Omega - q_\omega)}$$

Décision et pvalue (degré de signification)

Pour décider il suffit donc de considérer le seuil $f_{(n-q_\Omega);(q_\Omega-q_\omega)}(1-\alpha)$ lié au risque de première espèce

$$\alpha = P(\text{Décider } H_1 | H_0 \text{est vraie}) \quad (14)$$
$$1-\alpha = P(\text{Décider } H_0 | H_0 \text{est vraie}) \quad (15)$$
$$1-\alpha = P(F \leq f_{(n-q_\Omega);(q_\Omega-q_\omega)}(1-\alpha) | H_0 \text{est vraie}) \quad (16)$$
$$(17)$$

La pvalue est définie par

$$pvalue(F_{obs}) = P(F_{(n-q_\Omega);(q_\Omega-q_\omega)}(1-\alpha) > F_{obs} | H_0)$$

où F_{obs} est la réalisation de F.

Si $pvalue(F_{obs}) > \alpha$ alors on décide H_1.

Test sur tous les prédicteurs I

La première question à se poser est :

> *Est-ce qu'au moins une de mes variables explicative est utile ?*

On peut interpréter cette question comme un test de significativité du R^2.

$$\begin{cases} H_0 : & \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1 : & \exists j \in \{1, \ldots, p\} \ \beta_j \neq 0 \end{cases}$$

Test sur tous les prédicteurs II

Dans ce cas :

- $q_\omega = 1$
- $q_\Omega = p + 1$

Et la statistique de décision devient

$$F = \frac{SYY - RSS}{RSS} \frac{(n-p-1)}{(p)} \sim \mathcal{F}_{n-p-1;p}$$

qui est en fait le R^2 ajusté.

Test sur un unique prédicteur I

$$\begin{cases} H_0 : & \beta_j = 0 \\ H_1 : & \beta_j \neq 0 \end{cases}$$

L'hypothèse H_0 signifie que la variable x_j n'est pas liée à (n'apporte aucune information sur) Y.
Dans ce cas :

- $q_\omega = p$
- $q_\Omega = p + 1$

Test sur un unique prédicteur II

Une alternative équivalente consiste à faire un test de Student

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (X'X)^{-1}),$$

donc

$$\widehat{\beta}_j \sim \mathcal{N}(b_j, \sigma^2 v_j),$$

v_j désignant le terme diagonal (j, j) de la matrice $(X'X)^{-1}$. On peut encore écrire

$$\frac{\widehat{\beta}_j - \beta_j}{\sigma \sqrt{v_j}} \sim \mathcal{N}(0, 1).$$

Test sur un unique prédicteur III

En remarquant que $\widehat{\beta}_j$ et $\widehat{\sigma}^2$ sont indépendants (conséquence du Théorème Cochran encore), on a

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{v_j}} \sim \mathcal{T}_{n-p-1}.$$

Sous H_0, on a donc

$$\frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{v_j}} \sim \mathcal{T}_{n-p-1},$$

d'où la région critique du test, au niveau de signification α :

$$W : \frac{|\widehat{\beta}_j|}{\widehat{\sigma}\sqrt{v_j}} > t_{n-p-1;1-\alpha/2}.$$

Plan

Prédiction I

Lorsque les paramètres du modèle ont été estimés, et en supposant ce modèle valide, il est possible de l'utiliser pour *prédire* la valeur que prendra la variable Y pour de nouvelles valeurs des variables explicatives.
Posons $\boldsymbol{x}_0 = (1, x_{01}, \ldots, x_{0p})'$ le vecteur des variables d'entrée du modèle pour un nouvel individu. La sortie correspondante est

$$Y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

La quantité $\widehat{Y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$ fournit une prédiction non biaisée de Y_0, dans le sens où

$$\mathbb{E}(\widehat{Y}_0) = \boldsymbol{x}_0'\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0'\boldsymbol{\beta} = \mathbb{E}(Y_0).$$

Prédiction II

Il s'agit cependant d'une prédiction ponctuelle. Dans la pratique, il est important de donner une indication sur la « fiabilité »de la prédiction, ce que l'on peut faire en donnant :

- un intervalle de confiance sur $\mathbb{E}(Y_0)$ (un intervalle aléatoire contenant la constante $\mathbb{E}(Y_0)$ dans $100(1-\alpha)$ % des cas) ;
- un intervalle de prévision (un intervalle aléatoire contenant la v.a. Y_0 dans $100(1-\alpha)$ % des cas).

Prédiction III

Commençons par remarquer que $\widehat{Y}_0$ suit une loi normale. Il nous reste donc pour déterminer sa loi à calculer sa variance. On a

$$\text{var}(\widehat{Y}_0) = \boldsymbol{x}_0' \text{var}(\hat{\beta}) \boldsymbol{x}_0 = \boldsymbol{x}_0' [\sigma^2 (X'X)^{-1}] \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0' (X'X)^{-1} \boldsymbol{x}_0.$$

On a donc

$$\widehat{Y}_0 \sim \mathcal{N}(\boldsymbol{x}_0' \beta, \sigma^2 \boldsymbol{x}_0' (X'X)^{-1} \boldsymbol{x}_0).$$

Prédiction IV

On en déduit la fonction pivotale

$$\frac{\widehat{Y}_0 - \boldsymbol{x}_0'\beta}{\widehat{\sigma}\sqrt{\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}} \sim \mathcal{T}_{n-p-1},$$

qui conduit à l'intervalle de confiance suivant (au niveau de confiance $1-\alpha$) :

$$1-\alpha = P\left[\widehat{Y}_0 - t_{n-p-1;1-\alpha/2}\widehat{\sigma}\sqrt{\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0} < \mathbb{E}(Y_0) < \widehat{Y}_0 + t_{n-p-1;1-\alpha/2}\widehat{\sigma}\sqrt{\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}\right].$$

Prédiction V

Pour calculer un intervalle de prévision, on remarque que

$$Y_0 \sim \mathcal{N}(\boldsymbol{x}_0'\beta, \sigma^2)$$

d'où

$$\widehat{Y}_0 - Y_0 \sim \mathcal{N}(0, \sigma^2(1 + \boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0)).$$

On en déduit la fonction pivotale

$$\frac{\widehat{Y}_0 - Y_0}{\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}} \sim \mathcal{T}_{n-p-1},$$

Prédiction VI

et l'intervalle de prévision au niveau de confiance $1-\alpha$:

$$1-\alpha = P\left[\widehat{Y}_0 - t_{n-p-1;1-\alpha/2}\widehat{\sigma}\sqrt{1+\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0} < Y_0 < \widehat{Y}_0 + t_{n-p-1;1-\alpha/2}\widehat{\sigma}\sqrt{1+\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}\right].$$

On remarque que l'intervalle de prévision est plus large que l'intervalle de confiance.

Plan

Diagnostic I

L'objet du diagnostic est

- la vérification des hypothèses
  - linéarité
  - homoscédasticité
  - normalité
- la détection d'observations atypiques

Résidus I

L'examen des résidus joue un rôle fondamental. Il permet non seulement de vérifier empiriquement les hypothèses du modèle, mais également de détecter les observations atypiques (points aberrants) et de repérer les observations qui jouent un rôle important dans la détermination de la régression.
On appelle résidus bruts les quantités $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Afin de s'affranchir de facteurs d'échelle, il est utile de normaliser les résidus.
Soit $r_i = R_{ii}$ le terme diagonal (i, i) de la matrice $R = I - H$. On a donc

$$\text{var}(\widehat{\varepsilon}_i) = r_i \sigma^2$$

Résidus II

qui peut être estimé par $r_i \widehat{\sigma}^2$. On appelle *résidus studentisés* les quantités

$$s_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{r_i}}.$$

Afin de vérifier les hypothèses du modèle, on croise les résidus (bruts ou studentisés) avec les variables explicatives x_j et les prédictions $\widehat{Y}$

Distance de Cook (1977) I

Pour mettre en évidence ce type d'effet (influence « anormale »de certaines observations sur les résultats de la régression), on introduit les quantités suivantes, appelées *distances de Cook* :

$$D_i = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(-i)}\|^2}{(p+1)\widehat{\sigma}^2}$$

avec $\hat{\mathbf{Y}} = X\hat{\beta}$ et $\hat{\mathbf{Y}}_{(-i)} = X\hat{\beta}_{(-i)}$, $\hat{\beta}_{(-i)}$ étant l'estimation du vecteur des coefficients de régression obtenu en supprimant de l'ensemble d'apprentissage l'individu i (la ligne i de la matrice X et du vecteur $\mathbf{Y}$). La quantité D_i caractérise l'influence de l'observation i sur le résultat de la régression, une valeur élevée pouvant révéler une influence « anormale ».

Distance de Cook (1977) II

Remarquons que $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(-i)} = X(\hat{\beta} - \hat{\beta}_{(-i)})$, d'où

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})' X' X (\hat{\beta} - \hat{\beta}_{(-i)})}{(p+1)\widehat{\sigma}^2}$$

ce qui montre que D_i peut également s'interpréter comme le carré d'une distance entre les deux vecteurs $\hat{\beta}$ et $\hat{\beta}_{(-i)}$. On montre également que

$$D_i = \left[\frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{r_i}}\right]^2 \left[\frac{1 - r_i}{r_i}\right] \frac{1}{p+1},$$

où comme précédemment r_i est le terme diagnonal (i, i) de la matrice R. Il est donc inutile pour calculer les distances de Cook de refaire n fois les calculs de la régression.

Distance de Cook (1977) III

Quelle valeur de seuil choisir ? On peut montrer que la statistique de Cooks est une statistique de décision du test de Wald

$$H_0 : \beta = \beta_0$$

en supposant que la vrai valeur β_0 est la valeur estimée sans la ieme observation, et on peut utiliser une seuil $F_{p+1,n-p-1,1-\alpha}$

Plan

Qualité d'un modèle I

un modèle linéaire doit correspondre à un compromis :

- en augmentant le nombre de variables, on intègre de plus en plus d'information dans le modèle ;
- mais on augmente aussi la variance des estimations $\widehat{Y}_i$, car on augmente le nombre de paramètres à estimer.

En effet, on a

$$\text{var}(\widehat{\mathbf{Y}}) = \mathbb{E}[X(\hat{\beta}-\beta)(\hat{\beta}-\beta)'X'] = X\text{var}(\hat{\beta})X' = \sigma^2 X(X'X)^{-1}X' = \sigma^2 H.$$

La variance moyenne des $\widehat{Y}_i$ est donc

$$\frac{1}{n}\sum_{i=1}^{n}\text{var}(\widehat{Y}_i) = \sigma^2\frac{\text{trace}(H)}{n} = \sigma^2\frac{p+1}{n}.$$

Qualité d'un modèle II

On a donc intérêt à réduire p.
Pour cela, il faut choisir (1) un critère de qualité du modèle, et (2) une stratégie de sélection.

- Bayesian Information Criterion
- Akaike Information Criterion
- Cp de Mallows
- PRESS de Allen (identique à la validation croisée)

Akaike Information Criterion I

Un idéal pour le modèle serait de minimiser l'espérance de la distance entre le vecteur moyenne inconnu $\mu = \mathbf{X}\beta$ (précision de l'esimateur des moindres carrés). Si l'on considère la norme 2

$$\begin{aligned}
\|\mu - \mathbf{X}\hat{\beta}\|^2 &= \|\mathbf{y} - \epsilon - H\mathbf{y}\|^2 \\
&= \|\mathbf{y} - H\mathbf{y}\|^2 + \|\epsilon\|^2 - 2\epsilon^t(\mathbf{y} - H\mathbf{y}) \\
&= \|\mathbf{y} - H\mathbf{y}\|^2 + \|\epsilon\|^2 - 2\epsilon^t((\mu + \epsilon) - H(\mu + \epsilon)) \\
&= \|\mathbf{y} - H\mathbf{y}\|^2 - \|\epsilon\|^2 + 2\epsilon^t H\epsilon - 2\epsilon^t(\mu - H\mu)
\end{aligned}$$

Akaike Information Criterion II

et que l'on en prend l'espérance (risque quadratique)

- $\mathbb{E}[\|\epsilon\|^2] = n\sigma^2$
- $\mathbb{E}[2\epsilon^t H\epsilon] = 2\mathbb{E}[trace(\epsilon^t H\epsilon)] = 2trace(H\mathbb{E}[\epsilon^t\epsilon]) = 2trace(H)\sigma^2 = 2\sigma^2(p+1)$
- $\mathbb{E}[\epsilon^t(\boldsymbol{\mu} - H\boldsymbol{\mu})] = 0$

on obtient

$$\mathbb{E}[|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2] = \mathbb{E}[\|\mathbf{y} - H\mathbf{y}\|^2] - n\sigma^2 + 2trace(H)\sigma^2$$

On peut estimer cette quantité en remplaçant $\mathbb{E}[\|\mathbf{y} - H\mathbf{y}\|^2]$ par RSS/n et si l'on divise l'estimation par σ^2 on obtient AIC :

$$AIC = -2(\mathcal{L}(\hat{\beta}) + 2(p+1) + Cst$$

où $\mathcal{L}(\hat{\beta})$ est la log-vraisemblance du modèle.

Stratégie de séléction I

En ce qui concerne la stratégie de sélection de m variables parmi p variables initiales, on peut envisager, si p n'est pas trop grand, une recherche exhaustive (choix du meilleur sous-ensemble de variables parmi les p, au sens du critère retenu). Le nombre de sous-ensemble à tester est alors égal à $2^p - 1$, soit 31 pour $p = 5$, 1023 pour $p = 10$, 1048575 pour $p = 20$! En pratique, cette solution n'est donc faisable que pour une dizaine de variables initiales.

Stratégie de séléction II

Quand p est grand, il faut par conséquent avoir recours à une démarche heuristique sous-optimale. On utilise le plus souvent une procédure pas à pas consistant en l'élimination successive ou l'ajout successif de variables. On distingue notamment :

- la sélection ascendante : on ajoute incrémentalement des variables en maximisant à chaque fois le critère $\overline{R}^2$ (on cherche à chaque pas la variables qui fait décroître le plus la variance résiduelle) ;
- la sélection descendante : on commence avec les p variables, puis on retire à chaque pas la variable dont la suppression fait croître le moins la variance résiduelle.

Transformation

Pour améliorier la qualité du modèle, il est possible de transformer

- la variable à expliquer (Box-Cox utilise $t_\lambda(y)$ à la place de y)

$$t_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0 \\ \log y, & \text{sinon.} \end{cases}$$

- les variables explicatives. On ajoute souvent par exemple des puissances des variables explicatives, ou leur log si l'on soupçonne des effets multiplicatifs.

Contrastes I

Les variables qualitatives à K modalités sont codées par $K-1$ variables. Commençons par un exemple :

$$y_i = \mu + \alpha_j + \epsilon_i,$$

si y_i provient du groupe j.
Les paramètres μ et α_j sont définis de manière ambiguë. Les seules quantités clairement définies sont les

$$\mu_j = \mu + \alpha_j$$

mais l'on peut rajouter une constante à μ et la retrancher à chaque α_j.
La solution consiste à imposer une contrainte aux α_j.
Classiquement on poste $\sum_j \alpha_j = 0$.

Contrastes II

D'un point de vue matriciel, le codage initial ambiguë se traduit par une matrice de design

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

dont la dimension de l'espace colonnes n'est pas plein (chaque colonne peut être fabriquée à partir des 3 autres).
Une solution consiste à réduire le nombre de colonnes par combinaison linéaire en utilisant une matrice de contraste C_1

Contrastes III

$$X' = [1 : (X_1 C_1)]$$

où 1 est une colonne de 1, X_1 est la matrice composée des 3 dernières colonnes de X et

$$C_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

La matrice C_1 est une matrice de contraste. Elle remplace les 3 colonnes de X_1 par 2 colonnes indépendantes. De nombreuses solutions (matrices de contrastes existent), chaque contraste correspond à une interprétation.

Contrastes IV

- Contraste de traitement
- Contraste simple
- Contraste de tendance par polynôme orthogonaux
- ...

Sélection par régularisation

Le LASSO