SY09
La régression linéaire multiple

T. Denœux

1 Introduction

La régression linéaire multiple a pour but l'étude de la relation entre une variable à expliquer quantitative Y et p variables explicatives $x_1, \ldots, x_p$. Les variables explicatives sont supposées connues sans erreur et de nature non aléatoire (on les note par des lettres minuscules). Dans certains cas, il peut s'agir de variables contrôlées par l'utilisateur (conditions expérimentales, paramètres d'un process, etc.). La variable Y est, elle, une variable aléatoire. Sa loi de probabilité est supposée dépendre des x_j selon le modèle suivant :

$$Y = b_0 + b_1 x_1 + \ldots + b_p x_p + \varepsilon$$

avec $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. L'espérance de Y est donc fonction linéaire des entrées, tandis que la variance de Y, égale à σ^2 ne dépend pas des entrées (hypothèse d'homoscédasticité). L'équation précédente peut s'écrire vectoriellement :

$$Y = \mathbf{x}'\mathbf{b} + \varepsilon$$

avec $\mathbf{x} = (1, x_1, \ldots, x_p)' \in \mathbb{R}^{p+1}$ et $\mathbf{b} = (b_0, b_1, \ldots, b_p)' \in \mathbb{R}^{p+1}$.

Remarque 1 *La linéarité du modèle est essentiellement une linéarité* par rapport aux paramètres. *En effet, une relation linéaire par rapport aux paramètres mais non linéaire par rapport aux variables d'entrée peut toujours être linéarisée par changement de variable. Par exemple, si l'on a*

$$Y = b_0 + b_1 z^2 + b_2 \ln z + \varepsilon,$$

on peut toujours poser $x_1 = z^2$, $x_2 = \ln z$, *et retrouver le modèle linéaire précédent.*

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 :

$$Y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + \varepsilon_i \quad i = 1, n$$

avec $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

Matriciellement, ces n équations s'écrivent

$$\mathbf{Y} = X\mathbf{b} + \varepsilon$$

avec

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}$$

et

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

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

- estimation des paramètres $\mathbf{b}$ 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 $\mathbf{x}$;
- diagnostic de la régression (validation du modèle) ;
- sélection d'un ensemble de variables explicatives « pertinentes ».

2 Estimation des paramètres

2.1 Estimateur des moindres carrés de b

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

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

Théorème 1 *Le minimum de la fonction $S(\beta)$ est obtenu pour $\beta = \widehat{\mathbf{b}}$ avec*

$$\widehat{\mathbf{b}} = (X'X)^{-1} X'\mathbf{Y},$$

c'est-à-dire que l'on a

$$S(\widehat{\mathbf{b}}) = \min_{\beta} S(\beta).$$

$\widehat{\mathbf{b}}$ *est appelé* estimateur des moindres carrés *de* $\mathbf{b}$.

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 :

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

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

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

En supposant $X'X$ inversible, on obtient directement la solution de ce système :

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

Il reste à vérifier qu'il s'agit bien d'un maximum. Soit

$$\widehat{\mathbf{b}} = (X'X)^{-1}X'\mathbf{Y}$$

la solution trouvée précédemment, et

$$\widehat{\varepsilon} = \mathbf{Y} - X\widehat{\mathbf{b}}$$

le vecteur des écarts (appelés résidus). On a donc

$$S(\widehat{\mathbf{b}}) = \widehat{\varepsilon}'\widehat{\varepsilon}.$$

Soit $\widetilde{\mathbf{b}}$ une autre valeur de β. Le vecteur des écarts correspondant est

$$\widetilde{\varepsilon} = \mathbf{Y} - X\widetilde{\mathbf{b}} = (\mathbf{Y} - X\widehat{\mathbf{b}}) + (X\widehat{\mathbf{b}} - X\widetilde{\mathbf{b}}) = \widehat{\varepsilon} + X(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}}).$$

On a donc

$$\widetilde{\varepsilon}'\widetilde{\varepsilon} = \widehat{\varepsilon}'\widehat{\varepsilon} + 2(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})'X'\widehat{\varepsilon} + (\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})'X'X(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}}).$$

Le terme central du membre de gauche s'écrit

$$2(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})'X'\widehat{\varepsilon} = 2(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})'X'(\mathbf{Y} - X\widehat{\mathbf{b}}).$$

Or, d'après (2)

$$X'(\mathbf{Y} - X\widehat{\mathbf{b}}) = 0,$$

donc ce terme est nul et l'on a finalement

$$S(\widetilde{\mathbf{b}}) = S(\widehat{\mathbf{b}}) + (\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})'X'X(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}}).$$

Le dernier terme est une somme de carrés et ne peut donc être que positif ou nul. Par conséquent $S(\widetilde{\mathbf{b}}) \geq S(\widehat{\mathbf{b}})$. En conclusion, l'estimateur des moindres carrés de $\mathbf{b}$ est donc bien $\widehat{\mathbf{b}}$. □

On notera

$$\widehat{\mathbf{Y}} = X\widehat{\mathbf{b}} = X(X'X)^{-1}X'\mathbf{Y}$$

le vecteur des prédictions obtenu en remplaçant le paramètre $\mathbf{b}$ inconnu par son estimateur des moindres carrés $\widehat{\mathbf{b}}$.

Remarque 2 *On a supposé $X'X$ inversible, ce qui est le cas si la matrice X est de rang $p+1$. Si ce n'est pas le cas, c'est qu'une variable (une colonne de X) s'exprime comme combinaison linéaire des autres. Il suffit alors de supprimer la ou les variables redondantes.*

Remarque 3 *Si certaines variables sont très corrélées, la matrice $X'X$ est mal conditionnée est les calculs numériques peuvent être très imprécis. Une solution (appelée* ridge regression *en anglais) consiste à ajouter un terme sur la diagonale de $X'X$:*

$$\widehat{\mathbf{b}}_\lambda = (X'X + \lambda I)^{-1} X'\mathbf{Y}$$

où λ est une constante à déterminer. On montre que l'on améliore ainsi parfois les propriétés de l'estimateur. Une autre solution consiste a faire une analyse en composante principale préalable du tableau X, et à utiliser les composantes principales comme nouvelles variables (en supprimant celles correspondant à des valeurs propres nulles ou très faible). Cette technique porte le nom de régression sur composantes principales.

Remarque 4 *En pratique, il est inutile d'inverser la matrice $X'X$: il existe des algorithmes permettant de résoudre directement le système (2). En Matlab, on obtient directement la solution de* `Xb=Y` *par la commande* `b=X\Y`.

Remarque 5 *On a*

$$\widehat{\mathbf{Y}} = X\widehat{\mathbf{b}} = X(X'X)^{-1}X'\mathbf{Y} = P\mathbf{Y}$$

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

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

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

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

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

2.2 Propriétés de $\widehat{\mathbf{b}}$

Il est facile de calculer l'espérance et la variance de $\widehat{\mathbf{b}}$. On a les propriétés suivantes.

Théorème 2 *$\widehat{\mathbf{b}}$ est un estimateur sans biais de $\mathbf{b}$, et*

$$Var(\widehat{\mathbf{b}}) = \sigma^2(X'X)^{-1}.$$

Preuve : En effet,

$$\begin{aligned}\widehat{\mathbf{b}} &= (X'X)^{-1}X'(X\mathbf{b} + \varepsilon) \\ &= (X'X)^{-1}(X'X)\mathbf{b} + (X'X)^{-1}X'\varepsilon \\ &= \mathbf{b} + (X'X)^{-1}X'\varepsilon\end{aligned} \tag{3}$$

D'où

$$\mathbb{E}(\widehat{\mathbf{b}}) = \mathbf{b} + (X'X)^{-1}X'\mathbb{E}(\varepsilon) = \mathbf{b}.$$

Donc $\widehat{\mathbf{b}}$ est sans biais. Calculons sa variance :

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \mathbb{E}[(\widehat{\mathbf{b}} - \mathbf{b})(\widehat{\mathbf{b}} - \mathbf{b})'].$$

D'après ce qui précède,

$$\widehat{\mathbf{b}} - \mathbf{b} = (X'X)^{-1}X'\varepsilon, \tag{4}$$

d'où, compte-tenu du fait que $X'X$ est symétrique ($(X'X)' = X'X$) :

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \mathbb{E}[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}] = (X'X)^{-1}X'\mathbb{E}[\varepsilon\varepsilon']X(X'X)^{-1}.$$

Or, $\mathbb{E}[\varepsilon\varepsilon'] = \mathrm{Var}(\varepsilon) = \sigma^2 I_n$. Donc

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

□

Théorème 3 (Théorème de Gauss-Markov) *L'estimateur des moindres carrés $\widehat{\mathbf{b}}$ est optimal dans la classe $\mathcal{C}$ des estimateurs sans biais de $\mathbf{b}$ linéaires en $Y_1, \ldots, Y_n$.*

Cela signifie que, pour n'importe estimateur $\widetilde{\mathbf{b}}$ dans $\mathcal{C}$, la matrice $A = \mathrm{Var}(\widetilde{\mathbf{b}}) - \mathrm{Var}(\widehat{\mathbf{b}})$ est semi-définie positive : on a $\mathbf{x}'A\mathbf{x} \geq 0$ pour tout $\mathbf{x} \in \mathbb{R}^{p+1}$. Ce résultat est admis.

Théorème 4 *On a : $Cov(\widehat{\mathbf{b}}, \widehat{\varepsilon}) = 0$.*

Preuve : Par définition

$$\mathrm{Cov}(\widehat{\mathbf{b}}, \widehat{\varepsilon}) = \mathbb{E}[(\widehat{\mathbf{b}} - \mathbf{b})(\widehat{\varepsilon} - \mathbb{E}(\widehat{\varepsilon}))'].$$

Commençons par calculer $\mathbb{E}(\widehat{\varepsilon})$. On a

$$\widehat{\varepsilon} = R\mathbf{Y} = R(X\mathbf{b} + \varepsilon) = RX\mathbf{b} + R\epsilon.$$

Or, $RX = (I_n - X(X'X)^{-1}X')X = X - X = 0$.. Donc $\widehat{\varepsilon} = R\varepsilon$, et $\mathbb{E}(\widehat{\varepsilon}) = 0$. En utilisant l'équation (4) et sachant que $\widehat{\varepsilon} = R\varepsilon$, on a donc

$$\begin{aligned}\mathrm{Cov}(\widehat{\mathbf{b}}, \widehat{\varepsilon}) &= \mathbb{E}[(X'X)^{-1}X'\varepsilon\varepsilon'R'] \\ &= \sigma^2(X'X)^{-1}X'[I_n - X(X'X)^{-1}X'] = 0.\end{aligned}$$

□

2.3 Estimation de σ^2

Il reste à estimer la variance σ^2 des perturbations. Il est naturel pour cela de s'intéresser aux résidus $\widehat{\varepsilon}_i$, et plus particulièrement à la somme des carrés des résidus.

Théorème 5 *La variance des perturbations σ^2 est estimée sans biais par*

$$\widehat{\sigma}^2 = \frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2.$$

Preuve : On a

$$\widehat{\varepsilon}'\widehat{\varepsilon} = (R\varepsilon)'R\varepsilon = \varepsilon' R\varepsilon = \sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}\varepsilon_i\varepsilon_j,$$

d'où

$$\mathbb{E}(\widehat{\varepsilon}'\widehat{\varepsilon}) = \sigma^2 \sum_{i=1}^{n} R_{ii} = \sigma^2 \mathrm{Tr} R.$$

Or, on sait qu'on opérateur symétrique idempotent a toutes ses valeurs propres égales à 0 ou 1, donc sa trace est égale à son rang. Le rang de R est $n-p-1$ d'où le résultat. □

3 Analyse de la variance

3.1 Point de vue géométrique

Plaçons nous dans $\mathbb{R}^n$ et considérons les vecteurs

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

Le modèle linéaire s'écrit avec ces notations

$$\mathbf{Y} = b_0\mathbf{1} + \sum_{j=1}^{p} b_j\mathbf{x}_j + \varepsilon.$$

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 $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$. On cherche en effet

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

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

$$\widehat{\mathbf{Y}} = P\mathbf{Y},$$

P étant un opérateur de projection orthogonale.

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

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

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

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

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 :

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

3.2 Equation d'analyse de la variance

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

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

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

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

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 (S_{res}) est la variance des résidus, ou variance résiduelle.

Remarque 6 *La variance résiduelle est liée à l'estimateur sans biais $\widehat{\sigma}^2$ de σ^2 :*

$$\widehat{\sigma}^2 = \frac{n}{n-p-1}S_{res}.$$

Remarque 7 *A chacun des termes de l'équation d'analyse de la variance est associé un nombre de degrés de liberté (d.d.l.), égal au nombre de combinaisons linéaires des Y_i utilisées dans le calcul :*

- *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 $\widehat{Y}_i = \mathbf{x}_i'\widehat{\mathbf{b}}$ et $\overline{Y} = \overline{\mathbf{x}}'\widehat{\mathbf{b}}$. Par conséquent, le terme S_{reg} est fonction des paramètres $\widehat{b}_1, \ldots, \widehat{b}_p$ (le terme $\widehat{b}_0$ 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 $n-p-1$.*

TABLE 1 – Tableau d'analyse de la variance (SS : *sum of squares*; MS : *mean square*).

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

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 1).

3.3 Evaluation de la qualité de l'ajustement

On définit à partir de l'équation d'analyse de la variance le *coefficient de détermination*, égal à la proportion de la variance totale expliquée par la régression :

$$R^2 = \frac{S_{reg}}{S_{YY}} = 1 - \frac{S_{res}}{S_{YY}}.$$

Ce coefficient traduit la « qualité de l'ajustement », comme on le voit en considérant les deux situations extrêmes suivantes :

- Si les résidus sont nuls, on a $S_{res} = 0$ et $R^2 = 1$. Les n points $(\mathbf{x}_i, Y_i) \in \mathbb{R}^{p+1}$ sont alors situés dans l'hyperplan d'équation

  $$Y = \widehat{b}_0 + \widehat{b}_1 x_1 + \ldots + \widehat{b}_p x_p.$$

  Cela signifie que l'on peut retrouver sans erreur les Y_i à partir des $\mathbf{x}_i$, c'est-à-dire que toute la variation des Y_i est expliquée par les $\mathbf{x}_i$.
- Si les prédictions sont constantes ($\widehat{Y}_i = \overline{Y}, \forall i$), la variance expliquée est nulle et $R^2 = 0$. Dans ce cas, les $\mathbf{x}_i$ n'expliquent pas du tout la variation des Y_i.
- De manière générale, on a $0 \leq R^2 \leq 1$, et la valeur du R^2 s'interprète comme un « degré de liaison » entre les variables explicatives et la variable à expliquer.

Remarque 8 *Géométriquement, R^2 est égal au carré du cosinus de l'angle θ entre les vecteurs $\mathbf{Y} - \overline{\mathbf{Y}}$ et $\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}$: c'est donc le carré du coefficient de corrélation linéaire entre les Y_i et les $\widehat{Y}_i$.*

Remarque 9 *La définition du R^2 est telle que sa valeur augmente « mécaniquement » avec le nombre de variables explicatives : si on augmente la dimension du sous-espace $\mathcal{L}$, la distance de $\mathbf{Y}$ au sous-espace, et donc la variance résiduelle, ne peuvent que diminuer. Considérons l'exemple suivant. Supposons que l'on ait*

$p = 1$ variable explicative x, et $n = 4$ observations. Si l'on introduit 2 variables explicatives supplémentaires $x_2 = x^2$ et $x_3 = x^3$, le modèle devient

$$Y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \varepsilon.$$

Il est alors toujours possible de déterminer les 4 coefficients b_i de façon à avoir $\widehat{Y}_i = Y_i, \forall i$, et donc $R^2 = 1$, sans que cela ne traduise l'existence d'une relation plausible entre les variables x et y.

Pour obtenir un indicateur plus fiable de la qualité du modèle (et permettant la comparaison de plusieurs modèles ne possédant pas le même nombre de variables explicatives), on définit le R^2 ajusté comme suit :

$$\begin{aligned}\overline{R}^2 &= 1 - \frac{\frac{n}{n-p-1} S_{res}}{\frac{n}{n-1} S_{YY}} = 1 - \frac{n-1}{n-p-1} \frac{S_{res}}{S_{YY}} \\ &= 1 - \frac{n-1}{n-p-1}(1 - R^2) = \frac{n-1}{n-p-1} R^2 - \frac{p}{n-p-1}.\end{aligned}$$

On remarque que l'on a toujours $\overline{R}^2 \leq R^2$.

4 Tests de significativité

4.1 Loi des estimateurs sous hypothèse gausienne

Il est souvent intéressant de déterminer si les résultats de la régression (coefficients $\widehat{b}_j$ et R^2) sont dûs au hasard, ou s'ils traduisent l'existence d'une relation « significative » entre la variable à expliquer et les variables explicatives. Pour cela, il existe des tests de significativité, qui nécessitent une hypothèse supplémentaire : la normalité des perturbations :

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

En effet, sous cette hypothèse, il est possible de préciser la loi des estimateurs de $\mathbf{b}$ et de σ^2. En effet, on a

$$\widehat{\mathbf{b}} = [(X'X)^{-1}X']Y.$$

Or, $Y \sim \mathcal{N}(X\mathbf{b}, \sigma^2 I)$. Donc $\widehat{\mathbf{b}}$ suit une loi normale dans $\mathbb{R}^{p+1}$. Comme on a déjà calculé l'espérance et la variance de $\widehat{\mathbf{b}}$ dans le cas général, on a finalement

$$\widehat{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}, \sigma^2 (X'X)^{-1}).$$

Par ailleurs, on a

$$(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{1}{\sigma^2} \varepsilon' R \varepsilon,$$

car $\widehat{\varepsilon} = R\varepsilon$. Or, on a le théorème suivant (admis) :

Théorème 6 *Si $\mathbf{X}$ est un vecteur gaussien centré-réduit à composantes indépendantes, la forme quadratique $Q = \mathbf{X}' A \mathbf{X}$ suit une loi du χ^2 si et seulement si A est un projecteur orthogonal, c'est-à-dire si $A^2 = A$. Le rang de A est alors le degré de liberté du χ^2.*

La matrice R étant un projecteur orthogonal de rang $n-p-1$ (c'est la matrice de projection sur $\mathcal{L}^{\perp}$), on a

$$(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}. \tag{5}$$

4.2 Test de significativité d'un coefficient de régression

Soient les hypothèses suivantes :

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

L'hypothèse H_0 signifie que la variable x_j n'est pas liée à (n'apporte aucune information sur) Y. On a vu que

$$\widehat{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}, \sigma^2(X'X)^{-1}),$$

donc

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

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

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

D'après (5), et en remarquant que $\widehat{b}_j$ et $\widehat{\sigma}^2$ sont indépendants (conséquence du Théorème 4), on a

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

Sous H_0, on a donc

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

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

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

4.3 Test de significativité du R^2

Considérons maintenant les hypothèses :

$$\begin{cases} H_0 : & b_1 = b_2 = \ldots = b_p = 0 \\ H_1 : & \exists j \in \{1, \ldots, p\}\ b_j \neq 0 \end{cases}$$

L'hypothèse nulle signifie qu'il n'y a aucune liaison entre les variables explicatives et Y : le R^2 obtenu est donc non significatif, c'est-à-dire purement « accidentel ».

Reprenons l'équation de la variance :

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

On a vu que

$$\frac{nS_{res}}{\sigma^2} = (n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1},$$

quelque soit $\mathbf{b}$. Par ailleurs, sous H_0, les v.a. Y_i ont toutes la même loi, et donc nS_{YY}/σ^2 suit un χ^2_{n-1} comme variance empirique d'un échantillon de v.a. indépendantes de même loi. Enfin, S_{reg} ne dépend que de $\widehat{\mathbf{b}}$ et S_{res} ne dépend que de $\widehat{\varepsilon}$, donc ces 2 termes sont indépendants d'après le théorème 4. On en déduit que, sous l'hypothèse H_0, $nS_{reg}/\sigma^2 \sim \chi^2_p$.

Soit la statistique

$$F = \frac{nS_{reg}/\sigma^2 p}{nS_{res}/\sigma^2(n-p-1)} = \frac{S_{reg}/p}{S_{res}/(n-p-1)}.$$

D'après ce qui précède, F suit sous H_0 une loi de Fisher $F_{p,n-p-1}$. Sous H_1, le rapport de la variance expliquée à la variance résiduelle a tendance à prendre des valeurs plus élevées, d'où la région critique :

$$W : F > F_{p,n-p-1;1-\alpha}$$

Remarquons que F peut également s'exprimer en fonction de R^2. En effet,

$$F = \frac{n-p-1}{p}\frac{S_{reg}}{S_{res}} = \frac{n-p-1}{p}\frac{S_{reg}/S_{YY}}{1-S_{reg}/S_{YY}} = \frac{n-p-1}{p}\frac{R^2}{1-R^2},$$

d'où l'autre expression de la région critique :

$$W : \frac{n-p-1}{p}\frac{R^2}{1-R^2} > F_{p,n-p-1;1-\alpha}$$

Remarque 10 *Il peut arriver (rarement), que le test F soit significatif, et que chacun des tests t pour les hypothèses $H_0 : b_j = 0$ soit non significatif. L'inverse (test F non significatif mais certains coefficients significativement non nuls) n'est pas non plus impossible (mais encore plus rare).*

4.4 Test d'une sous-hypothèse linéaire

Il s'agit de tester l'hypothèse selon laquelle q coefficients sont nuls. Moyennant une permutation des indices, on peut toujours supposer que ce sont les q premiers, et écrire ainsi les hypothèses :

$$\begin{cases} H_0 : & b_1 = b_2 = \ldots = b_q = 0 \\ H_1 : & \exists j \in \{1, \ldots, q\}\ b_j \neq 0 \end{cases}$$

Pour résoudre ce test, il faut faire deux fois la régression, avec le modèle réduit (obtenu en ne prenant que les variables $q+1, \ldots, p$) et avec le modèle complet. Appelons S^0_{res} et S^1_{res} la variance résiduelle, respectivement dans le modèle réduit et dans le modèle complet, et considérons la statistique

$$F = \frac{S^0_{res} - S^1_{res}}{S^1_{res}}\frac{n-p-1}{q}.$$

En remarquant que $S_{res} = S_{YY}(1-R^2)$, F s'écrit encore

$$F = \frac{R^2_1 - R^2_0}{1-R^2_1}\frac{n-p-1}{q}$$

R_0^2 et R_1^2 désignant le coefficient de détermination dans le modèle réduit et dans le modèle complet. On peut montrer que, sous H_0,

$$F \sim F_{q,n-p-1},$$

d'où la région critique du test :

$$W : \frac{R_1^2 - R_0^2}{1 - R_1^2} \frac{n-p-1}{q} > F_{q,n-p-1;1-\alpha}$$

Ce test est très utile pour juger de la pertinence d'un ensemble de variables explicatives potentielles, en donnant un critère de significativité de l'augmentation du R^2 observée lorsqu'on complexifie le modèle initial.

5 Prédiction

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 $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0p})'$ le vecteur des variables d'entrée du modèle pour un nouvel individu. La sortie correspondante est

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

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

$$\mathbb{E}(\widehat{Y}_0) = \mathbf{x}_0'\mathbb{E}(\widehat{\mathbf{b}}) = \mathbf{x}_0'\mathbf{b} = \mathbb{E}(Y_0).$$

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).

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

$$\mathrm{Var}(\widehat{Y}_0) = \mathbf{x}_0'\mathrm{Var}(\widehat{\mathbf{b}})\mathbf{x}_0 = \mathbf{x}_0'[\sigma^2(X'X)^{-1}]\mathbf{x}_0 = \sigma^2\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0.$$

On a donc

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

On en déduit la fonction pivotale

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

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

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

Pour calculer un intervalle de prévision, on remarque que

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

d'où

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

On en déduit la fonction pivotale

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

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

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

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

6 Diagnostic de la régression

La phase de diagnostic de la régression consiste à vérifier (de manière plus ou moins subjective) que les hypothèses du modèle (linéarité de la relation entre les x_j et y, homoscédasticité, normalité des perturbations) sont adaptées aux données.

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. Pour cela on utilise le résultat suivant :

Proposition 1 *Var*$(\widehat{\varepsilon}) = \sigma^2 R$

Preuve : On a vu que $\widehat{\varepsilon} = R\varepsilon$ et $\mathbb{E}(\widehat{\varepsilon}) = 0$. On a donc

$$\mathrm{Var}(\widehat{\varepsilon}) = \mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}') = R\mathbb{E}(\varepsilon\varepsilon')R' = \sigma^2 RR' = \sigma^2 R.$$

□

Soit $r_i = R_{ii}$ le terme diagonal (i, i) de la matrice R. On a donc

$$\mathrm{Var}(\widehat{\varepsilon}_i) = r_i\sigma^2$$

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

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

Remarque 11 *Malgré l'appellation « résidus studentisés », les s_i ne suivent pas une loi de Student ($\widehat{\sigma}^2$ n'est pas indépendant de $\widehat{\varepsilon}_i$).*

Afin de vérifier les hypothèses du modèle, on croise les résidus (bruts ou studentisés) avec les variables explicatives x_j et les prédictions $\widehat{Y}$ (le croisement avec les Y_i a moins d'intérêt, car les résidus sont en général corrélés avec les Y_i). A l'examen de ces graphiques, on ne doit pas déceler de structure particulière (les points doivent être répartis de manière en apparence aléatoire à l'intérieur d'une bande de largeur à peu près constante). Sous hypothèse de normalité des perturbations, les résidus studentisés doivent par ailleurs être pratiquement tous compris entre -2 et +2. Si certaines hypothèses apparaissent comme non vérifiées, il faut modifier le modèle (transformation des Y_i, utilisation de modèles plus complexes qui sortent du cadre de ce cours).

L'examen des résidus n'est pas toujours suffisant pour détecter les points aberrants à cause de l'effet de levier : un point aberrant peut avoir une grande influence sur les coefficients de régression et avoir ainsi un résidu faible.

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* :

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

avec $\widehat{\mathbf{Y}} = X\widehat{\mathbf{b}}$ et $\widehat{\mathbf{Y}}_{(-i)} = X\widehat{\mathbf{b}}_{(-i)}$, $\widehat{\mathbf{b}}_{(-i)}$ 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 ».

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

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

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

7 Sélection des variables explicatives

Un dernier point important à considérer est le choix des variables explicatives. A partir d'un ensemble de variables connues susceptibles d'influer sur Y, il est possible de construire (à partir de transformations non linéaires : log, puissance, etc.) un nombre potentiellement très grand de variables explicatives x_i. En pratique, le nombre de variables à inclure dans le modèle 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

$$\mathrm{Var}(\widehat{\mathbf{Y}}) = \mathbb{E}[X(\widehat{\mathbf{b}} - \mathbf{b})(\widehat{\mathbf{b}} - \mathbf{b})'X'] = X\mathrm{Var}(\widehat{\mathbf{b}})X' = \sigma^2 X(X'X)^{-1}X' = \sigma^2 P.$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\widehat{Y}_i) = \sigma^2\frac{\mathrm{Tr}(P)}{n} = \sigma^2\frac{p+1}{n}.$$

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.

Le critère R^2 n'est pas un bon choix en général car il est monotone (on ne peut qu'augmenter le R^2 en ajoutant de nouvelle variables). Une alternative intéressante consiste à utiliser le R^2 ajusté, qui est un critère non monotone. Cela revient également à utiliser comme critère la variance résiduelle $\widehat{\sigma}^2$. En effet, on a

$$\widehat{\sigma}^2 = \frac{n}{n-p-1}S_{res}$$

et

$$\overline{R}^2 = 1 - \frac{\frac{n}{n-p-1}S_{res}}{\frac{n}{n-1}S_{YY}}$$

d'où l'on déduit

$$\widehat{\sigma}^2 = \frac{n}{n-1}(1-\overline{R}^2)S_{YY}.$$

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.

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.