Introduction à la régression multiple

Résumé

A la suite de la régression linéaire simple, cette vignette introduit le modèle linéaire multidimensionnel dans lequel une variable quantitative Y est expliquée, modélisée, par plusieurs variables quantitatives $X_j (j = 1, \ldots, p)$. Après avoir expliciter les hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle (moindres carrés), de prévision par intervalle de confiance, la signification des tests d'hypothèse sont discutées de même que les outils de diagnostics (graphe des résidus, colinéarité). Des développements complémentaires sont à rechercher dans une présentation plus complète du modèle linéaire.

Retour au plan du cours.

1 Introduction

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

2 Modèle

Une variable quantitative Y dite *à expliquer* (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $X^1, \ldots, X^p$ dites *explicatives* (ou encore de contrôle, endogènes, indépendantes, régresseurs).

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n ($n > p + 1$) de $\mathbb{R}^{(p+1)}$:

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i) \quad i = 1, \ldots, n.$$

L'écriture du *modèle linéaire* dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, X^1, \ldots, X^p\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de "1" . C'est-à-dire que les $(p + 1)$ variables aléatoires vérifient :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \quad i = 1, 2, \ldots, n$$

avec les hypothèses suivantes :

1. Les ε_i sont des termes d'erreur, non observés, indépendants et identiquement distribués ; $E(\varepsilon_i) = 0, Var(\varepsilon) = \sigma^2 \mathbf{I}$.
2. Les termes x^j sont supposés déterministes (facteurs contrôlés) **ou bien** l'erreur ε est indépendante de la distribution conjointe de $X^1, \ldots, X^p$. On écrit dans ce dernier cas que :

   $\mathbb{E}(Y|X^1, \ldots, X^p) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p$ et $\text{Var}(Y|X^1, \ldots, X^p) = \sigma^2$.
3. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.
4. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur ε ($\mathcal{N}(0, \sigma^2 \mathbf{I})$). Les ε_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice $\mathbf{X}(n \times (p + 1))$ de terme général x_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($x_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général y_i. En notant les vecteurs $\boldsymbol{\varepsilon} = [\varepsilon_1 \cdots \varepsilon_p]'$ et $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$, le modèle s'écrit matriciellement :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

3 Estimation

Conditionnellement à la connaissance des valeurs des X^j, les paramètres inconnus du modèle : le vecteur β et σ^2 (paramètre de nuisance), sont estimés par minimisation du critère des moindres carrés (M.C.) ou encore, en supposant (iv), par maximisation de la vraisemblance (M.V.). Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Attention, de façon abusive mais pour simplifier les notations, estimateurs et estimations des paramètres $\boldsymbol{\beta}$, c'est-à-dire la réalisation de ces estimateurs sur l'échantillon, sont notés de la même façon $\mathbf{b}$.

3.1 Estimation par M.C.

L'expression à minimiser sur $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - \cdots - \beta_p x_i^p)^2 &= \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

Par dérivation matricielle de la dernière équation on obtient les *"équations normales"* :

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

dont la solution correspond bien à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Des diagnostics de colinéarité et des aides au choix des variables sont explicités dans une présentation détaillée du modèle linéaire.

Alors, l'estimation des paramètres β_j est donnée par :

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

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{y}$ ont pour expression :

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

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est appelée "*hat matrix*"; elle met un chapeau à $\mathbf{y}$. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace Vect($\mathbf{X}$) engendré par les vecteurs colonnes de $\mathbf{X}$.

On note

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

le vecteur des résidus ; c'est la projection de $\mathbf{y}$ sur le sous-espace orthogonal de Vect($\mathbf{X}$) dans $\mathbb{R}^n$.

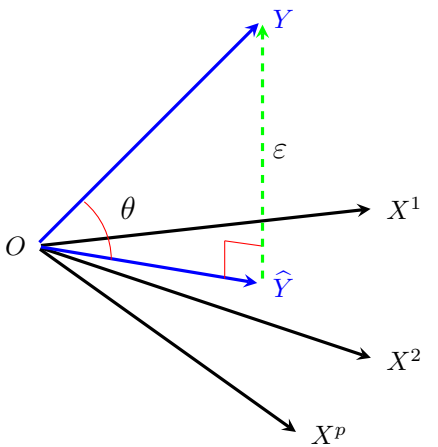

FIGURE 1 – Géométriquement, la régression est la projection $\widehat{Y}$ de Y sur l'espace vectoriel Vect$\{\mathbf{1}, X^1, \ldots, X^p\}$; de plus $R^2 = \cos^2(\theta)$.

3.2 Propriétés

Les estimateurs des M.C. $b_0, b_1, \ldots, b_p$ sont des estimateurs sans biais : $E(\mathbf{b}) = \boldsymbol{\beta}$, et, parmi les estimateurs sans biais fonctions linéaires des y_i, ils sont de variance minimum (propriété de Gauss-Markov) ; ils sont donc "BLUE" : *best linear unbiaised estimators*. Sous hypothèse de normalité, les estimateurs du M.V., qui coïncident avec ceux des moindres carrés, sont uniformément meilleurs ; ils sont efficaces c'est-à-dire que leur matrice de covariance atteint la borne inférieure de Cramer-Rao.

On montre que la matrice de covariance des estimateurs se met sous la forme

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

celle des prédicteurs est

$$E[(\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})(\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'] = \sigma^2\mathbf{H}$$

et celle des estimateurs des résidus est

$$E[(\mathbf{e} - \varepsilon)((\mathbf{e} - \varepsilon))'] = \sigma^2(\mathbf{I} - \mathbf{H})$$

tandis qu'un estimateur sans biais de σ^2 est fourni par :

$$s^2 = \frac{\|e\|^2}{n-p-1} = \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{n-p-1} = \frac{\text{SSE}}{n-p-1}.$$

Ainsi, les termes $s^2 h_i^i$ sont des estimations des variances des prédicteurs $\widehat{y}_i$.

3.3 Sommes des carrés

SSE est la somme des carrés des résidus (*sum of squared errors*),

$$\text{SSE} = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \|\mathbf{e}\|^2.$$

On définit également la somme totale des carrés (*total sum of squares*) par

$$\text{SST} = \|\mathbf{y} - \overline{y}\mathbf{1}\|^2 = \mathbf{y}'\mathbf{y} - n\overline{y}^2$$

et la somme des carrés de la régression (*regression sum of squares*) par

$$\text{SSR} = \|\widehat{\mathbf{y}} - \overline{y}\mathbf{1}\|^2 = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} - n\overline{y}^2 = \mathbf{y}'\mathbf{H}\mathbf{y} - n\overline{y}^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - n\overline{y}^2.$$

On vérifie alors : SST = SSR + SSE.

3.4 Coefficient de détermination

On appelle *coefficient de détermination* le rapport

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

qui est donc la part de variation de Y expliquée par le modèle de régression. Géométriquement, c'est un rapport de carrés de longueur de deux vecteurs. C'est donc le cosinus carré de l'angle entre ces vecteurs : $\mathbf{y}$ et sa projection $\widehat{\mathbf{y}}$ sur Vect$(\mathbf{X})$.

Attention, dans le cas extrême où $n = (p+1)$, c'est-à-dire si le nombre de variables explicatives est grand comparativement au nombre d'observations, $R^2 = 1$. Ou encore, il est géométriquement facile de voir que l'ajout de variables explicatives ne peut que faire croître le coefficient de détermination. Ce critère n'est qu'une indication de la *qualité d'ajustement* du modèle mais un R^2 proche de 1 n'est pas synonyme de bonne qualité de prévision. La quantité R est encore appelée *coefficient de corrélation multiple* entre Y et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{y}$ et sa prédiction (ou projection) $\widehat{\mathbf{y}}$.

4 Inférences dans le cas gaussien

En principe, l'hypothèse optionnelle (*iv*) de normalité des erreurs est nécessaire pour cette section. En pratique, des résultats asymptotiques, donc valides pour de grands échantillons, ainsi que des études de simulation, montrent que cette hypothèse n'est pas celle dont la violation est la plus pénalisante pour la fiabilité des modèles.

4.1 Inférence sur les coefficients

Pour chaque coefficient β_j on montre que la statistique

$$\frac{b_j - \beta_j}{\sigma_{b_j}}$$

où $\sigma_{b_j}^2$, variance de b_j est le j-ième terme diagonal de la matrice $s^2(\mathbf{X}'\mathbf{X})^{-1}$, suit une loi de Student à $(n-p-1)$ degrés de liberté. Cette statistique est donc utilisée pour tester une hypothèse $H_0 : \beta_j = a$ ou pour construire un intervalle de confiance de niveau $100(1-\alpha)\%$:

$$b_j \pm t_{\alpha/2;(n-p-1)}\sigma_{b_j}.$$

Attention, cette statistique concerne un coefficient et ne permet pas d'inférer conjointement sur d'autres coefficients car ils sont corrélés entre eux ; de plus elle dépend des absences ou présences des autres variables X^k dans le modèle. Par exemple, dans le cas particulier de deux variables X^1 et X^2 très corrélées, chaque variable, en l'absence de l'autre, peut apparaître avec un coefficient significativement différent de 0 ; mais, si les deux sont présentes dans le modèle, elles peuvent chacune apparaître avec des coefficients insignifiants.

De façon plus générale, si $\mathbf{c}$ désigne un vecteur non nul de $(p+1)$ constantes réelles, il est possible de tester la valeur d'une combinaison linéaire $\mathbf{c}'\mathbf{b}$ des paramètres en considérant l'hypothèse nulle $H_0 : \mathbf{c}'\mathbf{b} = a$; a connu. Sous H_0, la statistique

$$\frac{\mathbf{c}'\mathbf{b} - a}{(s^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})^{1/2}}$$

suit une loi de Student à $(n-p-1)$ degrés de liberté.

4.2 Inférence sur le modèle

Le modèle peut être testé globalement. Sous l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, la statistique

$$\frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} = \frac{\text{MSR}}{\text{MSE}}$$

suit une loi de Fisher avec p et $(n-p-1)$ degrés de liberté. Les résultats sont habituellement présentés dans un tableau *"d'analyse de la variance"* sous la forme suivante :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	*F*
Régression	p	SSR	MSR=SSR/p	MSR/MSE
Erreur	$n-p-1$	SSE	MSE=SSE/$(n-p-1)$	
Total	$n-1$	SST		

4.3 Inférence sur un modèle réduit

Le test précédent amène à rejeter H_0 dès que l'une des variables X^j est liée à Y. Il est donc d'un intérêt limité. Il est souvent plus utile de tester un modèle réduit c'est-à-dire dans lequel certains coefficients sont nuls (à l'exception du terme constant) contre le modèle complet avec toutes les variables. En ayant éventuellement réordonné les variables, on considère l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0, q < p$.

Notons respectivement SSR_q, SSE_q, R^2_q les sommes de carrés et le coefficient de détermination du modèle réduit à $(p-q)$ variables. Sous H_0, la statistique

$$\frac{(\text{SSR} - \text{SSR}_q)/q}{\text{SSE}/(n-p-1)} = \frac{(R^2 - R^2_q)/q}{(1-R^2)/(n-p-1)}$$

suit une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

Dans le cas particulier où $q = 1$ ($\beta_j = 0$), la F-statistique est alors le carré de la t-statistique de l'inférence sur un paramètre et conduit donc au même test.

4.4 Prévision par intervalle de confiance

Pour $\mathbf{x}_0$: $\widehat{y}_0 = b_0 + b_1 x^1_0 + \cdots + b_p x^p_0$

$$\widehat{y}_0 \pm t_{\alpha/2;(n-p-1)} s (1 + \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0)^{1/2}$$

5 Diagnostics

5.1 Résidus

Les mêmes diagnostics que dans le cas de la régression linéaire simple sont calculés en régression multiple : PRESS, distance de Cook et graphes des résidus. Ces graphes sont à regarder en tout premier pour apprécier les pré-requis (linéarité du modèle, homoscédasticité des résidus) et donc la validité du modèle. Toujours comme en régression simple, il importe de vérifier la normalité des résidus (droite de Henri) surtout dans le cas d'un échantillon restreint avec par exemple moins de 30 observations. Dans le cas d'un "grand échantillon", les propriétés asymptotiques des estimateurs corrigent un manque "raisonnable" de normalité des résidus : le modèle linéaire est dit *robuste* vis à vis de cette hypothèse.

5.2 Conditionnement

Attention Le point délicat de la régression multiple est généré par le calcul de la matrice inverse de $\mathbf{X}'\mathbf{X}$. Si cette matrice est *mal conditionnée* c'est-à-dire si son déterminant est proche de zéro, cela impacte directement la variance des estimateurs car des termes très grands apparaissent sur la diagonale de la matrice $\mathbf{H}$ dont dépendent directement les variances des estimations des paramètres comme celles des prévisions. Des indicateurs (facteurs d'inflation de la variance, rapports des valeurs propres) sont proposés pour alerter l'utilisateur d'un mauvais conditionnement mais des résultats mettant en évidence un nombre important de paramètres non significatifs (grandes p-valeurs des tests de Student) suffisent souvent pour détecter une telle situation : il y a sans doute "trop" de variables dans le modèle, certaines sont "presque" combinaisons linéaires des autres.

Cette remarque introduit toute la problématique du choix de modèle en régression lorsque l'objectif principal est de trouver un "meilleur" modèle de prévision : un modèle avec beaucoup de variables ajuste toujours (géométri-

quement) mieux les données mais court le risque d'un mauvais conditionnement, donc de plus grandes variances des estimations des paramètres et des prévisions. Cela affecte directement une estimation de l'erreur de prévision comme celle par exemple du PRESS obtenue par validation croisée. Des précisions sur les stratégies de recherche d'un meilleur modèle sont à lire dans une présentation détaillée du modèle linéaire.

La qualité de prévision d'un modèle ou plutôt celles de plusieurs modèles sont comparées en considérant une estimation de l'erreur quadratique de prévision :

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \widehat{y}_{(i)}\right)^2 = \sum_{i=1}^{n} \left(\frac{y_i - \widehat{y}_i}{1 - h_{ii}}\right)^2$$

6 Exemple

L'objectif de cette étude est de modéliser la note obtenue par des échantillons de fromage (Cheddar) lors de tests gustatifs opérés par un jury. Ce test concerne $n = 30$ échantillons de fromage dont la note moyenne doit être modélisée par $p = 3$ variables explicatives :

- `GoutM` : note moyenne de juges
- `Acetic` : log concentration en acide acétique
- `H2S` : log concentration en H2S
- `Lactic` : concentration en acide lactique

```
Analyse de régression : GoutM en fonction de Acetic; H2S; Lactic

Analyse de variance

Source            DL          SC          CM         F         P
Régression         3      4994,5      1664,8     16,22     0,000
Erreur résid      26      2668,4       102,6
Total             29      7662,9

L'équation de régression est
GoutM = - 28,9 + 0,33 Acetic + 3,91 H2S + 19,7 Lactic

Régresseur       Coef   Er-T coef          T         P
Constante     -28,88       19,74      -1,46     0,155
Acetic          0,328       4,460       0,07     0,942
H2S             3,912       1,248       3,13     0,004
Lactic         19,671       8,629       2,28     0,031

S = 10,13              R-carré = 65,2%      R-carré (ajust) = 61,2%
```

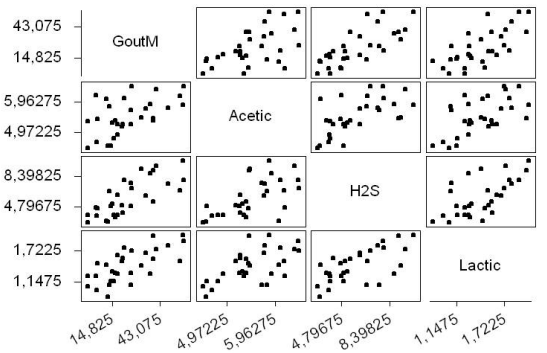

FIGURE 2 – matrice des nuages de points

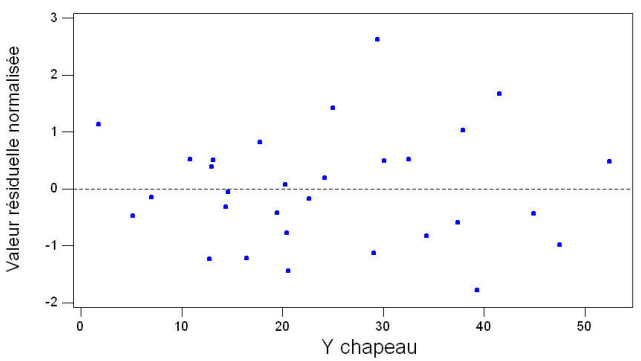

FIGURE 3 – Nuage des résidus normalisés sur les valeurs ajustées

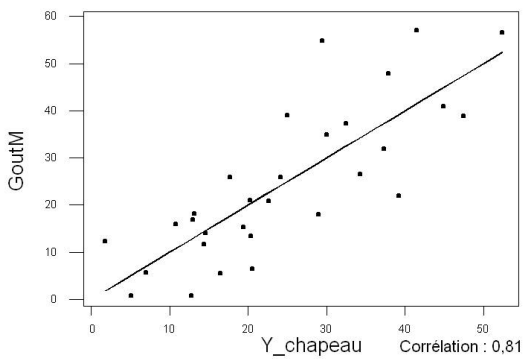

FIGURE 4 – Nuage des observations sur les valeurs ajustées