Régression

Cours de deuxième année de master

Bernard Delyon

28 septembre 2022

1. IRMAR, Université Rennes I, Campus de Beaulieu, 35042 Rennes cédex.

Table des matières

I

Introduction

I.1 L'objet de la régression.

Commençons par un exemple illustratif simple. Le botaniste Joseph Dalton Hooker a mesuré lors d'une expédition en 1849 la pression atmosphérique p_i et la température d'ébullition de l'eau t_i en divers endroits de l'Himalaya[1]. Selon les lois de la physique, $y_i = \ln(p_i)$ devrait être (en première approximation) proportionnel à t_i. On pose donc le modèle

$$y_i = \beta_1 + \beta_2 t_i + u_i. \tag{I.1}$$

u_i représente l'erreur de mesure (ou d'autres effets aléatoires), et explique que les points de la figure I.1 ne sont pas exactement alignés. Cette figure montre également la droite estimée par moindres carrés. On voit une très bonne adéquation. L'équation ci-dessus donne un modèle, qui si u_i est supposé gaussien centré devient le modèle paramétrique $y_i \sim \mathcal{N}(\beta_1 + \beta_2 t_i, \sigma^2)$, dont on verra l'intérêt plus tard. Le paramètre σ^2 représente la variance de l'écart des points à la droite (mesuré verticalement) et l'estimation de σ donne ici 0,2.

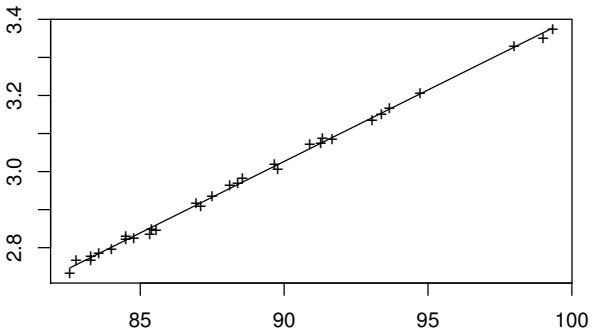


Figure I.1 – Logarithme de la pression mesurée en divers endroits de l'Himalaya en fonction de la température d'ébullition de l'eau.

Cet exemple illustre comment le modèle de régression tente d'expliquer au mieux une grandeur y (la **réponse**) en fonction d'autres grandeurs (**variables explicatives**, ou **régresseurs**, ou **facteurs**, la température dans l'exemple) en **démêlant ce qui est déterministe de ce qui est aléatoire** et en quantifiant ces deux aspects (par les β_i d'une part et σ^2 d'autre part).

1. En 1857 le physicien James David Forbes a fait la même expérience dans les Alpes, le but étant de pouvoir retrouver la pression atmosphérique à partir de la seule mesure de la température d'ébullition de l'eau (les baromètres étant fragiles et donc difficiles à transporter lors d'une expédition), ce qui permet ensuite d'en déduire l'altitude au travers d'une relation connue ; il rapporte dans un article ce double ensemble de données dont nous n'utilisons ici que la partie Himalayenne [78]

I.2 Exemples

I.2.1 Régression linéaire multiple : Production, travail et capital

On considère les variables, chacune concernant la totalité des États-Unis (i étant l'indice d'une année) :

- P_i : production
- K_i : capital (valeur des usines, etc.)
- T_i : travail fourni (basé sur un calcul du nombre total de travailleurs)

On cherche à expliquer P_i à l'aide des variables (K_i, T_i). Le modèle de Cobb et Douglas [38] est

$$P = \alpha_1 K^{\alpha_2} T^{\alpha_3}$$

ce qui suggère le modèle statistique

$$\log(P_i) = \log(\alpha_1) + \alpha_2 \log(K_i) + \alpha_3 \log(T_i) + u_i, \qquad E[u_i] = 0, \qquad E[u_i^2] = \sigma^2.$$

Les régresseurs sont donc ici $x_i = (1, \log(K_i), \log(T_i))$, la réponse est $y_i = \log(P_i)$ et les paramètres du modèle $\beta = (\log(\alpha_1), \alpha_2, \alpha_3)$. Le logarithme et les changements de variables ont permis de rendre le modèle linéaire (par rapport à β), ce qui, on le verra, est très avantageux pour l'analyse :

$$y_i = \beta_1 + \beta_2 \log(K_i) + \beta_3 \log(T_i) + u_i.$$

Cobb et Douglas disposaient du tableau suivant [2] sur $n = 24$ années et trouvent $\alpha_2 = 1/4$ et $\alpha_3 = 3/4$:

Année	P	K	T	Année	P	K	T	Année	P	K	T
1899	100	100	100	1907	151	176	138	1915	189	266	154
1900	101	107	105	1908	126	185	121	1916	225	298	182
1901	112	114	110	1909	155	198	140	1917	227	335	196
1902	122	122	118	1910	159	208	144	1918	223	366	200
1903	124	131	123	1911	153	216	145	1919	218	387	193
1904	122	138	116	1912	177	226	152	1920	231	407	193
1905	143	149	125	1913	184	236	154	1921	179	417	147
1906	152	163	133	1914	169	244	149	1922	240	431	161

I.2.2 Vers des modèles non-linéaires.

On observe des paires $(x_i, y_i)_{1 \leqslant i \leqslant n}$ où x_i où y_i est la concentration de produit actif dans un médicament au temps x_i après fabrication. Le modèle linéaire $y_i = \beta_1 + \beta_2 x_i + u_i$ est certainement inadéquat

Plusieurs modélisations peuvent être envisagées :

(a) Régression non-linéaire paramétrique : On part d'un modèle spécifique considéré comme réaliste

$$y_i = \beta_1 e^{-\beta_2 x_i} + u_i.$$

C'est l'analogue du précédent dans une situation non-linéaire.

(b) Régression polynomiale : On part cette fois-ci d'un modèle paramétrique abstrait

$$y_i = \sum_{j=0}^{J} \beta_j x_i^j + u_i.$$

où J est supposé connu. La linéarité en β de cette équation fait, on le verra, qu'on estime facilement les β_j par moindres carrés ; noter que ceci revient exactement à trouver le polynôme de degré J qui passe au plus près des points (x_i, y_i). On peut remplacer l'hypothèse $u_i \sim \mathcal{N}(0, \sigma^2)$ par l'hypothèse plus faible $E[u_i] = 0$ mais on entre alors dans un cadre semi-paramétrique.

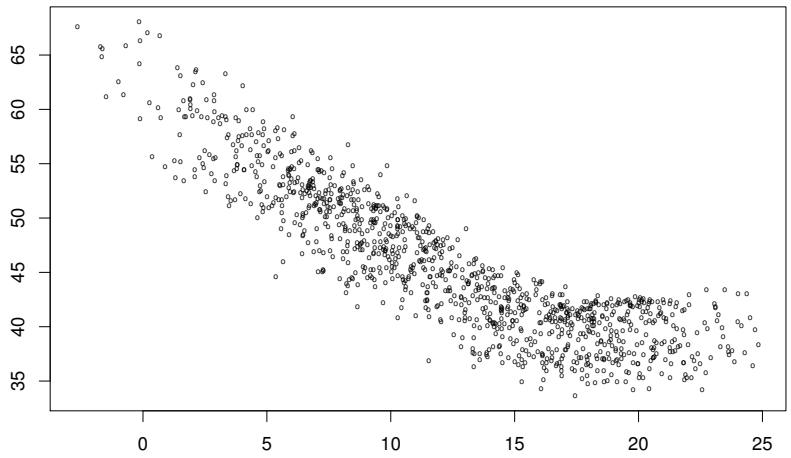
2. En réalité, la construction de ce tableau à partir des différentes données dont ils pouvaient disposer est en soi un travail énorme. Voir l'article.

(c) Régression non-paramétrique :

$$y_i = f(x_i) + u_i, \quad u_i = \mathcal{N}(0, \sigma^2).$$

Il s'agit d'estimer la fonction f et σ^2.

Un autre exemple. La figure suivante[3] représente la consommation d'électricité moyenne en France, à 2h du matin, en fonction de la température extérieure (moyenne sur les 24h précédentes). Les données sont sur 3 ans (1095 points). On pourrait être tenté de considérer ici un modèle linéaire par morceaux.



I.2.3 Modèle logistique : Credit scoring

Il s'agit pour une banque de mesurer le risque qu'elle prend à attribuer un crédit à un client.

La banque dispose de données sur ses anciens clients. Chaque client ayant demandé un crédit dans le passé est un individu et la réponse $y \in \{0, 1\}$ est une variable indiquant s'il y a eu un problème de remboursement. Le régresseur x est vecteur ligne contenant :

- des variables quantitatives : revenu, âge, dépôts, etc.
- des variables catégorielles : sexe, etc.

Le modèle logistique : y est une variable de Bernoulli $\mathcal{B}(1, p_x)$ (c-à-d $y = 1$ avec probabilité p_x) et p_x est de la forme

$$p_x = \frac{1}{1 + e^{-x\beta}}$$

où β est un vecteur colonne de paramètres caractérisant l'influence de chaque régresseur sur la réponse (de sorte que $x\beta$ est un produit scalaire). p_x représente le risque pris par la banque à autoriser un crédit au client ayant les régresseurs x.

I.2.4 Données longitudinales

On observe des variables

$$y_i(t_j) = F(t_j) + u_{ij}.$$

Par exemple $y_i(t_j)$ est la taille de l'enfant i au mois t_j. On se donne en général un modèle paramétrique particulier pour F, par exemple

$$F(t) = a + b \exp\left(-\exp\left(c - dt\right)\right).$$

3. Courtoisie de Vincent Lefieux, RTE.

Souvent un paramètre, disons b, dépendra des individus. Une méthode simple pour prendre cette dépendance en compte sera de rassembler les caractéristiques d'intérêt de l'individu i (végétarien/non-végétarien, taille des parents, etc.) dans un vecteur (ligne) x_i et présupposer une relation linéaire, ce qui donne finalement le modèle

$$y_i(t_j) = a + (x_i\beta)\exp(-\exp(c - dt_j)) + u_{ij}$$

($x_i\beta$ est un produit scalaire) dont les paramètres sont (a, c, d, β).

I.3 Méthode générale et objectifs de la régression.

On peut voir la régression comme le cadre le plus simple pour la modélisation paramétrique des suites de variables aléatoires indépendantes non-stationnaires[4]. En pratique, les applications essentielles sont les suivantes (on illustre ici par l'exemple rudimentaire où y est le taux de fréquentation du médecin et x contient l'âge et le sexe de l'individu) :

- Détermination des **facteurs significatifs** : L'âge a-t-il une influence significative sur le taux de fréquentation du médecin ? (c.-à-d. : le coefficient β_i de l'âge est-il nul ?)
- **Prédiction/simulation** (des réponses connaissant les régresseurs et β) : Combien de médecins faut-il pour une ville de pyramide des âges donnée ?
- **Détection de changement** (du paramètre β) : Une modification du ticket modérateur a-t-il provoqué un changement significatif dans le comportement des patients ? Ce changement est-il le même chez les hommes et chez les femmes ?

La méthode passe, comme on vient de le voir, par la mise en place d'un modèle plus ou moins réaliste sur lequel il est bon d'avoir du recul : on peut le considérer comme un (pâle) reflet de la réalité mais il est généralement plus prudent d'y voir simplement un **instrument de mesure** qui permettra de quantifier certains phénomènes tout en restant maître de ce que l'on calcule.

I.4 Exercices

Exercice 1. On dispose de deux qualités de papier. Le papier de type 1 a un poids β_1 et le papier de type 2 a un poids β_2 (grammes par feuille). On reçoit n paquets. Le i-ième paquet contient p_i feuilles du type 1 et q_i feuilles du type 2. On pèse successivement les paquets sur une balance ; le poids mesuré du i-ième paquet est m_i. On admet que les erreurs de la balance sont $\mathcal{N}(0, \sigma^2)$. Écrire le modèle linéaire correspondant à ces données.

Exercice 2. (Modèle gravitaire) On suppose que le nombre de personnes de la ville i allant travailler à la ville j suit en gros le modèle idéal suivant

$$N_{ij} = k d_{ij}^{-\alpha} P_i A_j$$

où P_i est a population de la ville i, A_i sa capacité d'accueil et d_{ij} la distance entre les villes. k et α sont des paramètres inconnus. Proposer un modèle de régression linéaire pour des données basées sur I villes $\{d_{ij}, N_{ij}, P_i, A_i, 1 \leqslant i, j \leqslant I\}$.

Exercice 3. Un individu pris au hasard a un temps de réaction à un certain stimulus qui suit la loi $\mathcal{N}(\mu, \sigma^2)$, μ et σ^2 sont connus. Après absorption d'une quantité x d'alcool, ce temps se trouve multiplié par $1 + \beta x$. On s'intéresse à l'estimation de β à partir de données (x_i, t_i), où t_i est le temps de réaction. Expliciter la loi de t_i. Peut-on poser ce problème comme un problème de régression linéaire simple ? Qu'en est-il si maintenant β est connu et μ est le paramètre ? Proposer alors un estimateur de μ.

4. Certains modèles de régression, comme les modèles mixtes, prennent toutefois en compte des phénomènes de dépendance.

II

Régression linéaire multiple

II.1 Introduction

II.1.1 Les données

Les données consistent en des variables observées y_i (réponses) et des variables explicatives (ou régresseurs) x_i, $i = 1, \dots n$, chaque paire (y_i, x_i) représentant une expérience (un individu). On les arrange dans un tableau de la façon suivante :

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

x_i est donc un vecteur ligne contenant les p variables explicatives. On convient généralement de mettre le régresseur constant, s'il est présent, dans la première colonne.

On présuppose l'existence d'une relation du type $y_i \simeq \langle x_i, \beta^* \rangle = x_i \beta^*$ pour un certain vecteur (colonne) β^*, soit $y \simeq X\beta^*$, ce qui conduit au modèle de régression linéaire

$$y = X\beta^* + u$$

où $u = (u_1, \dots u_n)$ est un vecteur de bruit (variables aléatoires) modélisant l'inadéquation des mesures au modèle.

Le but de la régression linéaire est l'estimation de β^* et la validation du modèle. La valeur de l'estimée obtenue sera notée $\widehat{\beta}$. Ceci se fera en minimisant en β une certaine norme du vecteur $y - X\beta$.

II.1.2 L'hypothèse de rang plein

Il est clair que si X n'est pas de rang colonnes plein c'est-à-dire s'il existe v tel que $Xv = 0$ (une combinaison linéaire des colonnes est nulle) alors pour tout β

$$X\beta = X(\beta + v).$$

Ceci implique que pour tout estimateur $\widehat{\beta}$, l'estimateur $\widehat{\beta} + v$ explique aussi bien les données. Par conséquent on ne pourra pas estimer β^* à moins de faire des hypothèses supplémentaires. Une autre façon de le voir est de remarquer que comme $Xv = 0$ une de colonne de X (et sans doute chacune) est fonction linéaire des autres, et par conséquent une des variables étant fonction linéaire des autres est inutile.

Pour cette raison X sera généralement supposée rang colonnes plein (ce qui signifie aussi que $X^T X$ est inversible, puisque $Xv = 0$ est sans solution).

II.1.3 Le régresseur constant

Il est très généralement présent mais pas toujours. Toutefois, l'essentiel des résultats énoncés dans la suite (tests de Fisher) reste valide sans cette hypothèse.

II.2 Moindres carrés ordinaires

II.2.1 Modèle statistique et interprétation

Modèle. *On suppose l'existence d'un vecteur β^*, de $\sigma_* > 0$ et de variables aléatoires u_i tels que*

$$\begin{aligned} y &= X\beta^* + u, \\ E[u] &= 0, \\ E[uu^T] &= \sigma_*^2 I. \end{aligned}$$

En d'autres termes, pour chaque i :

$$\begin{aligned} y_i &= x_i\beta^* + u_i \\ E[u_i] &= 0 \\ Var(u_i) &= \sigma_*^2 \qquad \text{(homoscedasticité)} \\ E[u_i u_j] &= 0, \quad j \neq i \quad \text{(décorrélation des bruits).} \end{aligned}$$

Noter que ce modèle n'est pas complètement spécifié puisque les lois des u_i ne sont pas précisées. On est pour l'instant dans une situation semi-paramétrique.

II.2.2 Estimation de β^* et σ_*^2

1 - Définition

Soit $SS(\beta)$ (Sum of Squares) la somme des carrés des erreurs de prédiction

$$SS(\beta) = \|y - X\beta\|^2 = \sum (y_i - x_i\beta)^2.$$

L'estimateur de β^* aux moindres carrés ordinaires (Ordinary Least Squares, OLS) est

$$\widehat{\beta} = \arg\min_{\beta} SS(\beta).$$

C'est l'estimateur de β^* au maximum de vraisemblance sous l'hypothèse de normalité de u.

Ceci correspond, dans la figure I.1, à minimiser la somme des carrés des "distances" des points à la droite *mesurées verticalement*; il pourrait sembler plus logique de minimiser la somme des carrés des vraies distances, mais cet autre estimateur $\check{\beta}(X, y)$ est plus compliqué à calculer et n'est pas invariant par changement d'échelle au sens où $\check{\beta}(X, ty) \neq t\widehat{\beta}(X, y)$ (car une homothétie en y modifie complètement le calcul des distances; cf. § II.4).

2 - Proposition

On a les propriétés :

- $\widehat{\beta} = (X^T X)^{-1} X^T y$
- $\widehat{\beta} = \beta^* + (X^T X)^{-1} X^T u$
- $\widehat{\beta}$ est sans biais : $E[\widehat{\beta}] = \beta^*$
- $Var(\widehat{\beta}) = \sigma_*^2 (X^T X)^{-1}$

La démonstration est laissée en exercice (ex. 2 p. 21). Il est intéressant de noter que si la variable $X_{.j}$ est décorrélée des autres, et $X_{.1}$ vaut 1, alors $\widehat{\beta}_j$ est insensible au retrait d'autres variables ; et $\widehat{\beta}_k$, $k \neq j$ est insensible au retrait de $X_{.j}$.

3 - Proposition

Soit

$$RSS = SS(\widehat{\beta}) = \|y - X\widehat{\beta}\|^2$$

(Residual Sum of Squares) ; alors l'estimateur suivant de σ_*^2 est sans biais :

$$\widehat{\sigma}^2 = RSS/(n-p).$$

La démonstration est présentée à la suite de la proposition 5.

4 - Définition

- Vecteur des valeurs ajustées (fitted values) : $\widehat{y} = X\widehat{\beta}$
- Vecteur des résidus (residuals) : $\widehat{u} = y - \widehat{y}$
- Erreur standard de $\widehat{\beta}_j$ est $\widehat{\sigma}(\widehat{\beta}_j)$ défini par : $\widehat{\sigma}(\widehat{\beta}_j)^2 = \widehat{\sigma}^2[(X^TX)^{-1}]_{jj}$.

Exemple. Reprenons le modèle de Cobb-Douglass du paragraphe I.2.1 avec les données de leur étude de 1928. On trouve $\widehat{\beta}_2 = 0,23$ et $\widehat{\beta}_3 = 0,81$. L'écart entre $\widehat{\beta}_2 + \widehat{\beta}_3$ et 1, n'est en fait pas significatif, ce qu'on peut vérifier en utilisant les résultats du chapitre suivant. L'erreur standard de $\widehat{\beta}_2 + \widehat{\beta}_3$ peut être obtenue et l'on trouve 0,09. Les résultats en sortie de logiciel sont présentés ainsi :

	Estimate $\widehat{\beta}_j$	Std. Error $\widehat{\sigma}(\widehat{\beta}_j)$
(Intercept)	-0.177	0.434
log(capital)	0.233	0.063
log(travail)	0.807	0.145

« Intercept » désigne le coefficient de la constante, $\beta_1 = \ln \alpha_1$. Il est estimé ici à $\simeq 0$ en raison de la normalisation des données.

II.2.3 Propriétés géométriques élémentaires

Dans toute la suite, pour tout vecteur z, $\bar{z}$ désignera la moyenne de ses coordonnées

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i.$$

5 - Proposition

Soit $H = X(X^TX)^{-1}X^T$, $K = I - H$, et $\mathcal{X}$ le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de X ; alors

- H est le projecteur orthogonal sur $\mathcal{X}$; K est le projecteur orthogonal sur $\mathcal{X}^\perp$.
- $\widehat{y} = Hy, \quad \widehat{u} = Ky = Ku, \quad \widehat{y} \perp \widehat{u}.$

Et s'il y a une colonne constante dans la matrice X :

- $\bar{\widehat{u}} = 0 \quad$ car $\quad \widehat{u} \perp \mathbf{1}$
- $\|y - \bar{y}\mathbf{1}\|^2 = \|y - \widehat{y}\|^2 + \|\widehat{y} - \bar{y}\mathbf{1}\|^2$
  $\sum_i (y_i - \bar{y})^2 = \sum_i \widehat{u}_i^2 + \sum_i (\widehat{y}_i - \bar{y})^2$
  TSS = RSS + ESS
  Var. Totale = Var. Résiduelle + Var. Expliquée

La démonstration est laissée en exercice. Cette décomposition de la variance correspond à l'idée présentée dans l'introduction de séparer le déterministe de l'aléatoire.

Pour la démonstration de la proposition 3, on a : $\widehat{\sigma}^2 = \|\widehat{u}\|^2/(n-p) = u^T K u/(n-p)$ d'où,

$$E[\widehat{\sigma}^2] = \sigma^2 Tr(K)/(n-p) = \sigma^2.$$

II.2.4 Le coefficient de corrélation multiple R

On introduit ici le coefficient R^2 qui mesure l'adéquation géométrique du modèle aux données. On suppose ici que X contient une colonne constante.

6 - Définition

R est la corrélation empirique entre les données et les valeurs prédites

$$R = \frac{\sum_i (y_i - \bar{y})(\widehat{y}_i - \bar{y})}{(\sum_i (y_i - \bar{y})^2)^{1/2} (\sum_i (\widehat{y}_i - \bar{y})^2)^{1/2}}.$$

R^2 est appelé coefficient de détermination, ou encore la proportion de variance expliquée.

L'interprétation la plus simple est de voir R comme une mesure de corrélation entre les variables explicatives (globalement) et les réponses. Plus R est proche de 1, plus le modèle représente bien les données. Par exemple sur la figure I.1 on a $R^2 = 0,998$.

7 - Proposition

On a

- $0 \leqslant R \leqslant 1, \quad R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$,
- $R = 1 \iff \widehat{y} = y$
- $R = 0 \iff \widehat{\beta} = (\bar{y}, 0, \ldots, 0)$.

Démonstration.

$$R = \frac{\langle \widehat{y} - \bar{y}\mathbf{1}, y - \bar{y}\mathbf{1} \rangle}{\|\widehat{y} - \bar{y}\mathbf{1}\| \, \|y - \bar{y}\mathbf{1}\|} = \frac{\langle \widehat{y} - \bar{y}\mathbf{1}, y - \widehat{y} + \widehat{y} - \bar{y}\mathbf{1} \rangle}{\|\widehat{y} - \bar{y}\mathbf{1}\| \, \|y - \bar{y}\mathbf{1}\|} = \frac{\|\widehat{y} - \bar{y}\mathbf{1}\|}{\|y - \bar{y}\mathbf{1}\|} = \frac{\sqrt{ESS}}{\sqrt{TSS}}.$$

Si $R = 1$ alors $RSS = 0$, $y = \widehat{y}$. Si $R = 0$ alors $\widehat{y} = \bar{y}\mathbf{1}$ et donc $X\widehat{\beta} = X(\bar{y}, 0, \ldots, 0)^T$ d'où $\widehat{\beta} = (\bar{y}, 0, \ldots, 0)^T$ car X est de rang plein. ∎

Attention, le R^2 ne dit pas tout sur la qualité du modèle ; par exemple, les figures II.1 ont même valeur de R. R^2 doit être considéré comme une donnée descriptive, intéressante en soi, et pratique pour comparer des modèles sur les mêmes données, mais il ne peut être considéré comme une note absolue : même si le modèle est valide, R^2 est une variable aléatoire dont la distribution (de même que celle de $\widehat{\beta}$) peut dépendre fortement de la répartition des régresseurs (à moins que $\beta_j^* = 0$, $j > 1$). Noter également que l'ajout d'un régresseur fera toujours augmenter R^2, même si le β_j^* correspondant est nul.

Le R^2 sera utilisé plus tard dans le cadre bien précis du test de Fisher de nullité de β^*, p. 49.

Le R^2 ajusté vaut $R^2_{aj} = 1 - \widehat{\sigma}^2/(TSS/(n-1))$; c'est un rapport d'estimées débiaisées de variances. Il tente d'exprimer le pourcentage de variance expliquée qui serait mesuré sur de nouveaux échantillons. L'ajout d'un régresseur ne fait pas nécessairement augmenter R^2_{aj}.

Exemple. Dans le cas du modèle de Cobb-Douglas, on trouve $TSS = 2,3$, $ESS = 1,6$ et $RSS = 0,7$. On a donc $R^2 = 0,7$. On dit que *le travail et le capital investi expliquent 70% de la variabilité de la production* (en fait son logarithme).

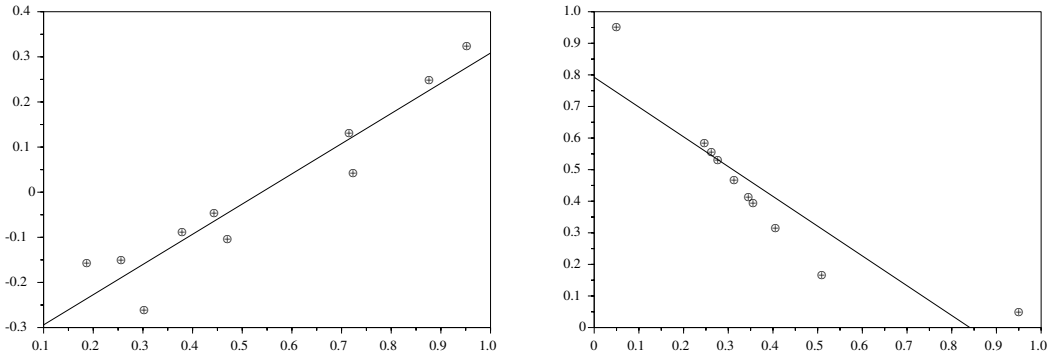
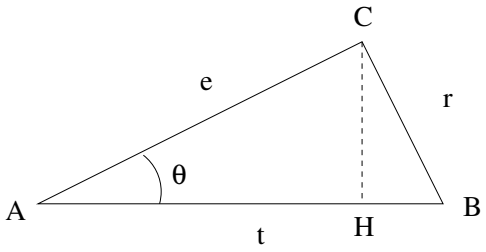


FIGURE II.1 – Points (x_i, y_i) et la droite de régression. Deux exemples de même R^2.

Pourquoi raisonner sur les carrés pour juger des contributions et non pas sur les valeurs absolues ? Voici un dessin qui peut le justifier :



Le côté AC a pour longueur $e = \sqrt{ESS}$, et de même avec $r = \sqrt{RSS}$ et $t = \sqrt{TSS}$. Il est raisonnable de juger que la contribution des variables explicatives est donnée par AH/AB et celle des résidus par BH/AB. On a bien

$$\frac{AH}{AB} = \frac{AH}{AC}\frac{AC}{AB} = \cos(\theta)^2 = \frac{e^2}{t^2}.$$

Le rapport de carrés est donc en fait également un rapport de deux longueurs.

II.2.5 Effet de la suppression d'un individu. Effet levier

Le coefficient $h_i = H_{ii} = x_i(X^TX)^{-1}x_i^T$ (leverage) mesure l'éloignement du i-ième individu x_i des autres ; plus précisément (cf. exercice 13 p. 22) :

8 - PROPOSITION

On a

- $0 < h_i \leqslant 1$
- $h_i = 1 \iff span(x_j,\ j \neq i)$ est de dimension $p-1$
- $\lim_{\|x_i\|\to\infty} h_i = 1$

Une valeur élevée de h_i indique que le vecteur x_i est *isolé* soit parce que sa norme est élevée, soit parce qu'il est le seul présent dans une direction donnée. Il sera donc influent dans l'estimation de $\widehat{\beta}$ (effet levier), et on dit que h_i est un indice d'influence du régresseur x_i. Tout ceci sera précisé au § III.3.

On l'obtient sous `R` avec la commande `h=lm.influence(mod)$hat`.

Si les données sont bien réparties, les h_i sont à peu près égaux à p/n (on sait que $\sum h_i = \text{trace}(H) = \text{rang}(H) = p$).

Soit $X_{(i)}$ la matrice X dont on a retiré la i-ième ligne x_i et $y_{(i)}$ le vecteur y dont on a retiré le i-ième coefficient. Soient $\widehat{\beta}_{(i)}$ et $\widehat{\sigma}_{(i)}$ les estimées aux moindres carrés de β^* et σ_* basées sur $X_{(i)}$ et $y_{(i)}$. Alors (cf. exercice 13 p. 22) :

9 - Théorème

Après suppression de la i-ième observation, les estimateurs aux moindres carrés des paramètres deviennent

$$\widehat{\beta}_{(i)} = \widehat{\beta} - (X^T X)^{-1} x_i^T \frac{\widehat{u}_i}{1-h_i} \tag{II.1}$$

$$(n-p-1)\widehat{\sigma}^2_{(i)} = (n-p)\widehat{\sigma}^2 - \frac{\widehat{u}_i^2}{1-h_i} \tag{II.2}$$

II.2.6 Effet de l'ajout d'un régresseur et coefficient de corrélation partielle

On part du modèle précédent

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

puis on rajoute un régresseur, c'est-à-dire une colonne à X

$$X' = (X, \xi).$$

On se propose de trouver une formule permettant de passer directement de $\widehat{y}$ à $\widehat{y}'$, pour pouvoir ensuite calculer l'évolution du coefficient de détermination. On a besoin du lemme suivant :

10 - Lemme

Soient A et B deux sous-espaces vectoriels orthogonaux de $\mathbb{R}^n$, alors en notant P_A, P_B et $P_{A,B}$ les projecteurs orthogonaux sur A, B et sur $A \oplus B$, on a

$$P_{A,B} = P_A + P_B.$$

Démonstration. Soit x un vecteur montrons que $P_A x + P_B x$ est bien $P_{A,B}x$. Le vecteur $P_A x + P_B x$ appartient bien à $A \oplus B$ et de plus $x - P_A x - P_B x$ est orthogonal à A (car $x - P_A x$ et $P_B x$ le sont) et de la même façon à B ; donc $P_A x + P_B x$ coïncide avec $P_{A,B}x$. ∎

Notons $\xi_\perp = (\xi - P_{\mathcal{X}}\xi)/\|\xi - P_{\mathcal{X}}\xi\|$ la composante de ξ orthogonale à $\mathcal{X}$ normalisée.

$$\widehat{y}' = H'y = P_{\mathcal{X},\xi} y = P_{\mathcal{X},\xi_\perp} y = P_{\mathcal{X}} y + P_{\xi_\perp} y = \widehat{y} + \langle \xi_\perp, y\rangle \xi_\perp = \widehat{y} + \langle \xi_\perp, \widehat{u}\rangle \xi_\perp \tag{II.3}$$

car $\widehat{y} \perp \xi_\perp$; notons que le dernier terme est la prédiction du résidu par $\xi_\perp$. Le nouveau vecteur de résidus est

$$\widehat{u}' = \widehat{u} - \langle \xi_\perp, \widehat{u}\rangle \xi_\perp$$

et par application du théorème de Pythagore, comme $\widehat{u}' \perp \xi_\perp$, on a

$$\|\widehat{u}\|^2 = \|\widehat{u}'\|^2 + \langle \xi_\perp, \widehat{u}\rangle^2.$$

Donc finalement, le nouveau coefficient R' satisfait

$$1 - R'^2 = \frac{\|\widehat{u}'\|^2}{\|y - \bar{y}1_n\|^2} = \frac{\|\widehat{u}\|^2}{\|y - \bar{y}1_n\|^2} \frac{\|\widehat{u}'\|^2}{\|\widehat{u}\|^2} = (1-R^2)(1-\rho^2)$$

avec

$$\rho^2 = \frac{\|\widehat{u}\|^2 - \|\widehat{u}'\|^2}{\|\widehat{u}\|^2} = \frac{\langle \xi_\perp, \widehat{u}\rangle^2}{\|\widehat{u}\|^2}. \tag{II.4}$$

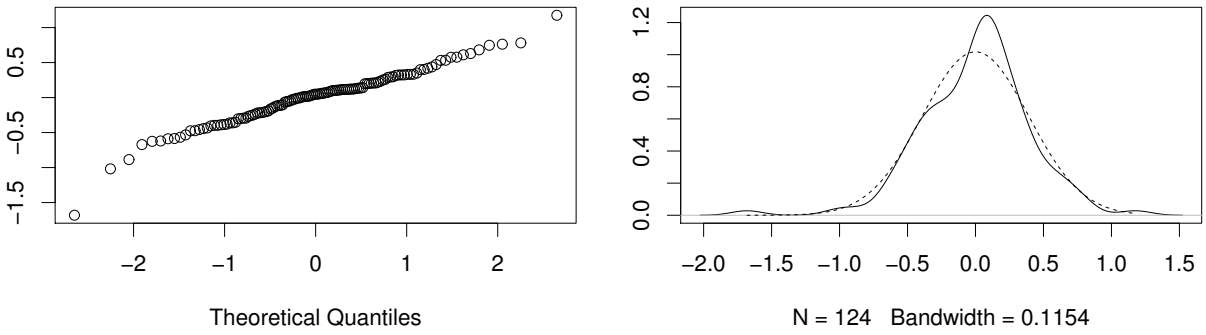
$\rho = \cos(\widehat{\xi_\perp, \widehat{u}})$ est appelé coefficient de corrélation partielle de y et ξ sachant x, car c'est la corrélation des variables dont on a retranché la projection sur $\mathcal{X}$ (alors que d'habitude on se contente de les recentrer). C'est l'analogue du coefficient R où cette fois on cherche à prédire au mieux $\widehat{u}$ à l'aide du régresseur $\xi_\perp$. Sa valeur absolue mesure l'apport du nouveau régresseur pour la qualité de la prédiction.

On vérifie sans difficulté que *ρ est du signe du coefficient β affecté à ξ* (multiplier (II.3) par $\xi_\perp^T$). Ce signe n'est pas forcément celui de la corrélation usuelle comme dans l'exemple suivant : Dans la prédiction du prix des voitures en fonction des certaines variables et du taux de CO_2, on peut remarquer que ce taux a une corrélation positive avec le prix (les grosses voitures...) mais un coefficient négatif dans la régression, pour une raison évidente, donc une corrélation partielle négative.

II.2.7 Aspects pratiques. Représentation graphiques exploratoires

Histogramme des résidus. Droite de Henry (QQ-plot). Test de Shapiro. Il s'agit de vérifier l'hypothèse de normalité. La droite de Henry s'approxime raisonnablement de la façon suivante : ordonner les résidus standardisés $\widehat{u}_i/\widehat{\sigma}$ puis les tracer en fonction de $Q(i/(n+1))$ (quantile de la loi supposée, ici la gaussienne) ; si la distribution des $\widehat{u}_i$ est normale, on doit trouver des points approximativement alignés.

Ne pas oublier toutefois que la distribution des $\widehat{u}_i/\widehat{\sigma}$ n'est pas exactement $\mathcal{N}(0,1)$ (surtout pour n petit, cf. § III.3). On représente ici le QQ-plot correspondant aux données de la figure II.2, avec le deuxième modèle ; la coïncidence semble assez bonne sauf pour trois individus. En revanche un estimateur de densité supperposé avec la gaussienne correspondante donne une adéquation peu probante, ce que le test de Shapiro confirme (p-value de 0.006).



Représentation résidus/valeurs ajustées. C'est une représentation des $\widehat{u}_i$ en fonction des $\widehat{y}_i$. L'estimation fait que ces deux variables sont empiriquement décorrélées ; toutefois la représentation peut faire apparaître une dépendance évidente. Elle peut provenir par exemple d'une non-linéarité de la relation liant y_i à x_i, ou plus souvent d'une variance non-constante des u_i.

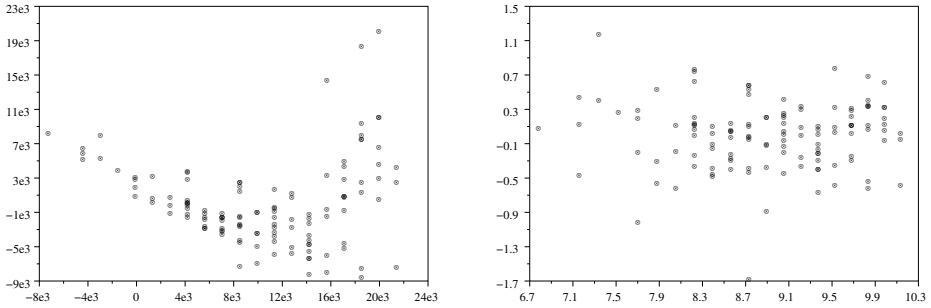


FIGURE II.2 – Prix de voitures d'occasion (réponse) en fonction de l'âge (régresseur) (Source : OzDASL). Le modèle est $p_i = \beta_1 + \beta_2 a_i + u_i$. Sur la première figure est tracé le résidu en fonction du prix prédit. La seconde est similaire mais avec les nouvelles variables $\log(p)$ et $\log(a)$.

Représentation réponses/régresseurs. On trace y_i en fonction de x_{ij} à j fixé. C'est une méthode

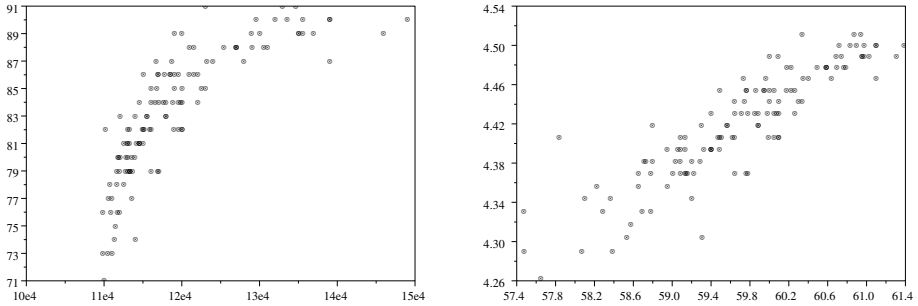


FIGURE II.3 – Mêmes données que la figure II.2. On trace les résidus partiels, $y - \widehat{\beta}_1$, en fonction du prix pour les deux modèles. La figure réponse/régresseur est ici la même à un décalage vertical près.

rustique pour détecter une dépendance non-linéaire entre un régresseur donné et la réponse. Ceci peut conduire à **ajouter aux régresseurs** des fonctions de ces derniers (par exemple x_{i2}^2, voir aussi l'introduction du temps dans l'exemple du § II.5.3), quitte à les éliminer plus tard lors des tests.

Représentation des résidus partiels. Ce tracé sert à illustrer l'influence du j-ième régresseur x^j (j-ième vecteur colonne de X), pour vérifier par exemple l'hypothèse de linéarité. Précisons dès maintenant que cette méthode doit être employée avec précautions car *s'il y a de fortes corrélations entre variables, ou si le modèle est trop inexact, ce tracé peut donner des résultats très mauvais, bien pires que la représentation réponses/régresseurs.* L'idée est de tracer le résidu obtenu sans x^j

$$z = y - X\widehat{\beta} + x^j\widehat{\beta}_j = \widehat{u} + x^j\widehat{\beta}_j \tag{II.5}$$

en fonction de x^j. Ce tracé tente d'illustrer la dépendance de y en x^j, la contribution des autres variables ayant été réduite au maximum ; l'apparition d'une structure particulière (autre qu'une droite) peut remettre en cause l'hypothèse de linéarité.

Mallows [60] recommande d'ajouter $(x^j)^2$ en variable explicative afin de mieux prendre en compte la non-linéarité potentielle de x^j ; ceci fait un terme en plus dans le membre de droite de (II.5).

On les obtient sous R par la commande `residuals(...,type="partial")`, ou bien en utilisant la commande `crp` de la bibliothèque `car`.

Étude théorique. Si l'on note e_j le j-ième vecteur de la base canonique de $\mathbb{R}^p$, alors

$$z = y - X(I - e_j e_j^T)\widehat{\beta} = (I - XP_j(X^TX)^{-1}X^T)y = Qy$$

où $P_j = I - e_j e_j^T$ est le projecteur orthogonal sur l'orthogonal de e_j. On vérifie sans peine que Q est le projecteur oblique de noyau x^k, $k \neq j$, sur l'espace contenant x^j et les vecteurs orthogonaux aux colonnes de X. L'effet de Q est donc de « nettoyer » la contribution linéaire des autres régresseurs en conservant celle de x^j.

Noter que cette méthode est tout-à-fait différente du choix $z = y - X^{(j)}\widehat{\beta}^{(j)}$ où l'exposant j signale la suppression du j-ième régresseur, car ici la matrice Q ne préserve pas x^j.

Étude des régresseurs. On pourra faire une analyse de X (corrélations entre variables, présence de sous-groupes d'individus, ACP de X, etc.). On verra au chapitre suivant qu'une dépendance entre variables explicatives peut fortement troubler l'analyse.

Échelles. Il est souvent utile de remettre les données sur une échelle correcte. Voir l'exemple de la figure II.2.

Une situation classique quand y est positif est d'observer graphiquement que la variance est en gros proportionnelle à y^2, ce qui indique que c'est plutôt $\log y$ qui suit un modèle homoscédastique ; en effet on voit facilement que si $\log y = x\beta + u$ et $\sigma_u \ll 1$, alors $y \sim e^{x\beta}(1+u)$, ce qui fait une variance en

y^2. Cette transformation est souvent utilisée lorsque la variance augmente avec y, et $y > 0$. On appelle parfois cela la **stabilisation de variance**.

Plus généralement on utilise les transformations de Box et Cox[1] : $\frac{y^\lambda - 1}{\lambda}$ avec $0 \leqslant \lambda \leqslant 1$. On choisit alors traditionnellement λ en maximisant le R^2.

II.2.8 Traitement des variables catégorielles

En pratique on a souvent affaire à des variables catégorielles (qualitatives). La méthode la plus courante pour prendre en compte une telle variable dans une étude statistique est de la convertir en plusieurs variables à valeurs 0 ou 1 ; par exemple si la classe d'âge d'un individu a trois modalités, J, A, V, on remplacera la variable explicative $x \in \{J, A, V\}$ par un vecteur $x' = (1_{x=A}, 1_{x=J}, 1_{x=V})$, ce qui permet ensuite d'employer des méthodes numériques. Malheureusement, on voit que toute composante de x' est fonction des deux autres ce qui fait que cette méthode est en général mathématiquement inutilisable telle quelle, aussi bien en analyse de données qu'en régression (on va voir plus bas qu'elle conduit à une matrice X de rang déficient) ; le procédé habituel consiste à ôter arbitrairement une modalité, si bien que x' devient $x' = (1_{x=A}, 1_{x=J})$. Ceci conduit à une situation particulièrement embrouillante que l'on détaille ici (même si les logiciels habituels font automatiquement ces transformations).

Si l'on cherche à expliquer y_i (p.ex. la productivité à l'hectare) à l'aide d'une variable catégorielle z_i prenant disons 3 valeurs a, b, c (p.ex. « pas d'engrais », « engrais1 », « engrais2 »), on peut être tenté de fabriquer les variables

$$x_{i1} = 1, \quad x_{i2} = 1_{z_i=a}, \quad x_{i3} = 1_{z_i=b}, \quad x_{i4} = 1_{z_i=c}$$

et d'utiliser le formalisme précédent. Toutefois, si l'on fait cela, le tableau X ne sera pas de rang plein car on a pour tout i $x_{i1} - x_{i2} - x_{i3} - x_{i4} = 0$, ou encore

$$X \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = 0.$$

Ceci signifie simplement la présence d'une variable en trop. Pour résoudre ce problème il suffit d'éliminer une des 4 variables. Ainsi on considérera le modèle

$$y_i = \beta_1 1_{z_i=a} + \beta_2 1_{z_i=b} + \beta_3 1_{z_i=c} + u_i. \tag{II.6}$$

ou bien

$$y_i = \beta'_1 + \beta'_2 1_{z_i=a} + \beta'_3 1_{z_i=b} + u_i \tag{II.7}$$

Ces deux paramétrisations sont équivalentes ($\beta'_1 = \beta_3$, $\beta'_2 = \beta_1 - \beta_3$, $\beta'_3 = \beta_2 - \beta_3$) et donnent, si les données sont bien rangées, une matrice X de la forme

$$X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \quad \text{et} \quad X' = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{pmatrix}.$$

1. Pour une discussion approfondie, voir [29].

S'il n'y a qu'une variable catégorielle, la représentation (II.6) est la plus naturelle ; en revanche, s'il y en a plusieurs il est plus simple de se contenter de retrancher à chaque fois une modalité :

$$y_i = \beta_1 + \beta_2 1_{z_i=a} + \beta_3 1_{z_i=b} + \beta_4 1_{t_i=n} + u_i, \qquad z_i \in \{a,b,c\}, \qquad t_i \in \{n,s\} \tag{II.8}$$

Ces complications viennent fondamentalement du fait que ce dernier modèle (modèle additif) n'est en réalité pas naturel du tout (le bon modèle étant donné par (II.10)). On verra que l'avantage du modèle sans interaction (II.8) est d'avoir moins de paramètres ; il a ici $1 + (3-1) + (2-1) = 4$ paramètres.

Si des variables quantitatives sont présentes, il suffit bien entendu de les ajouter au tableau X ; on peut soit considérer que leur influence est indépendante de la (ou des) variable catégorielle :

$$y_i = \beta_1 + \beta_2 1_{z_i=a} + \beta_3 1_{z_i=b} + \beta_4 x_i + u_i$$

soit qu'elle en dépend (modèle avec interactions)

$$y_i = \beta_1 + \beta_2 1_{z_i=a} + \beta_3 1_{z_i=b} + \beta_4 x_i + \beta_5 1_{z_i=a} x_i + \beta_6 1_{z_i=b} x_i + u_i. \tag{II.9}$$

P.ex. x est la pluviosité et dans ce dernier modèle l'augmentation de productivité en présence de pluie peut dépendre de l'engrais utilisé ; cette dépendance est reflétée par la valeur de β_5 ou β_6 en comparaison de β_4.

On a donc ici 6 régresseurs, ce qui correspond aux contributions de la constante (1), de x (1), de z $(3-1=2)$ et de l'interaction $((3-1)\times 1 = 2)$. De même, pour un modèle à deux variables catégorielles avec p et q modalités, le modèle sans interaction aura $1 + (p-1) + (q-1)$ régresseurs, et le modèle avec interactions en aura $pq = 1 + (p-1) + (q-1) + (p-1)\times(q-1)$, avec des interactions du type $1_{z=a}1_{z'=a'}$, $1_{z=b}1_{z'=a'}\ldots$ Ce mode de calcul s'étend à un nombre arbitraire de variables. Noter que comme dans le cas des équations (II.6,II.7), on a les modélisations équivalentes

$$\begin{aligned} y &= \beta_1 1_{z=a,t=n} + \beta_2 1_{z=b,t=n} + \beta_3 1_{z=c,t=n} + \beta_4 1_{z=a,t=s} + \beta_5 1_{z=b,t=s} + \beta_6 1_{z=c,t=s} + u \\ y &= \beta'_1 + \beta'_2 1_{z=a} + \beta'_3 1_{z=b} + \beta'_4 1_{t=n} + \beta'_5 1_{z=a,t=n} + \beta'_6 1_{z=b,t=n} + u. \end{aligned} \tag{II.10}$$

Il faut bien voir que dans le cas d'un modèle complet avec toutes les interactions entre variables catégorielles, comme ci-dessus, le décompte des paramètres ne pose aucun problème, il suffit de calculer toutes les possibilités, sans le régresseur constant (formulation (II.10)) ; ceci est également valide dans le cas où se mêlent variables catégorielles et quantitatives, par exemple le modèle (II.9) se réécrit plus simplement

$$y_i = \beta_1 1_{z_i=a} + \beta_2 1_{z_i=b} + \beta_3 1_{z_i=c} + \beta_4 1_{z_i=a} x_i + \beta_5 1_{z_i=b} x_i + \beta_6 1_{z_i=c} x_i + u_i$$

avec $3+3=6$ paramètres ; si l'on ajoute t, on a alors 12 paramètres. La gymnastique de décompte proposée plus haut n'a par conséquent d'intérêt que si l'on considère des modèles où toutes les interactions ne sont pas prises en compte comme (II.8).

Exemple. On observe la prise de poids de rats nourris avec quatre régimes différents correspondant à deux sources de protéines possibles (bœuf ou céréales) en deux doses possibles (faible ou élevée) citesnedecor,hand. Chacune des combinaisons des deux facteurs est testée sur 10 individus tous différents ; il y a donc 40 observations de prise de poids en tout. Le tableau des coefficients estimés en sortie de R se présente comme suit

	coef.
(Intercept)	100
DoseFaible	-20.8
ProtéineCéréale	14.4
DoseFaible : ProtéineCéréale	18.8

et $\widehat{\sigma}$ est donné à 15. Le modèle avec interactions estimé se réécrit

$$poids = 100 - 20,8\ 1_{D=f} - 14,1\ 1_{P=c} + 18,8\ 1_{D=f,P=c} + 15\ \mathcal{N}(0,1).$$

La prise de poids consécutive à un régime de bœuf à dose faible est de moyenne 79,2 avec un écart-type de 15.

II.2.9 Exercices

Exercice 1. Préciser la matrice X de l'exercice 1 p. 10. À quoi correspondrait l'ajout du régresseur constant ?

Exercice 2. Démontrer les propositions 2 et 5.
Indication : Pour la vérification de ce que $\widehat{\beta}$ minimise bien SS, on admettra que $H = X(X^TX)^{-1}X^T$ est le projecteur sur $\mathcal{X}$ (l'espace engendré par les colonnes de X ; cf. proposition 5), on montrera que pour tout β, $y - X\beta = H(\widehat{\beta} - \beta) + (I - H)y$. Ceci permettra, avec l'aide du théorème de Pythagore, d'exprimer $SS(\beta)$ sous une forme sous laquelle il devient évident que $\widehat{\beta}$ minimise SS. Un autre procédé consiste à noter que $\partial_{\beta_j} SS(\beta) = \sum_i 2x_{ij}(y_i - x_i\beta) = 2(X^T(y - X\beta))_j$, d'où l'équation $X^T(y - X\beta) = 0$.

Exercice 3. On est dans la situation de l'exercice 1 p. 10 avec $(p_1, p_2, p_3) = (50, 40, 60)$ et $q_i = 100 - p_i$. Expliciter sa valeur de l'estimateur OLS de β si $y_1 = y_2 = 1$, et $y_3 = 2$; qu'observe-t-on ? Peut-on ajouter le régresseur constant ?

Exercice 4. Démontrer que s'il n'y a qu'un régresseur en dehors de la constante (i.e. $p = 2$), alors R est la corrélation empirique entre x et y.

Exercice 5. On considère le modèle

$$y_i = bx_i + u_i, \qquad E[u_i] = 0, \qquad E[u_i^2] = \sigma^2, \qquad E[u_iu_j] = 0$$

où x_i est scalaire. Expliciter l'estimateur des moindres carrés $\widehat{b}$. Soit l'estimateur $\check{b} = \sum y_i / \sum x_i$. Comparer ces deux estimateurs en calculant leur biais et leur variance (On vérifiera que la propriété BLUE (exercice 12) s'applique bien : la variance de $\widehat{b}$ est inférieure à celle de $\check{b}$). Sous quelle condition les variances sont-elles égales ?

Exercice 6. Calculer la covariance entre $\widehat{u}$ et $\widehat{\beta}$ (on pourra utiliser la relation $\widehat{u} = Ku$).

Exercice 7. On fait une régression de y sur deux variables explicatives x et z, c-à-d $X = (\mathbf{1}, x, z)$; il y a en tout n individus. On a obtenu le résultat suivant :

$$X^TX = \begin{pmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

1. Que vaut n ? Que vaut le coefficient de corrélation linéaire empirique entre x et z ? (Indication : penser à l'interprétation de chaque entrée de X^TX en fonction des colonnes de X).

La régression linéaire fournit les résultats :

$$y = -1 + 3x + 4z + \widehat{u}, \quad RSS = 3.$$

2. Que vaut la moyenne empirique $\bar{y}$ (on utilisera la matrice X^TX) ?
3. Calculer $\|\widehat{y}\|^2$; justifier que $\|\widehat{y} - \bar{y}\mathbf{1}\|^2 = \|\widehat{y}\|^2 - \|\bar{y}\mathbf{1}\|^2$; en déduire ESS, TSS et le coefficient de détermination R^2.

On s'intéresse au modèle privé du régresseur z :

$$y = X_0\beta_0 + u_0, \quad X_0 = (\mathbf{1}_n, x).$$

4. Calculer numériquement X_0^Ty (commencer par calculer X^Ty) ; en déduire $\widehat{\beta}_0$.
5. Calculer $\|\widehat{y}_0\|^2$. Démonter que $\|\widehat{u}_0\|^2 + \|\widehat{y}_0\|^2 = \|\widehat{u}\|^2 + \|\widehat{y}\|^2$. En déduire la norme de $\widehat{u}_0$.
6. Calculer le coefficient de corrélation partielle entre z et y sachant x.

Exercice 8. Vérifier les propriétés de la matrice Q du § II.2.7.

Exercice 9. Erreur en prédiction. C_p Mallows.

1. Vérifier que $E[\widehat{u}_i^2] = \sigma_*^2(1-h_i)$. On pourra commencer par démontrer la formule suivante : $\widehat{u}_i^2 = e_i^T K u_i u_i^T K^T e_i$ où e_i est le i-ième vecteur de la base canonique.
   En déduire que $\widehat{\sigma}^2$ est sans biais.
2. Soit un nouvel individu (y', x') satisfaisant les hypothèses du modèle, vérifier que

$$E[(y' - x'\widehat{\beta})^2] = \sigma_*^2(1 + x'(X^TX)^{-1}x'^T).$$

3. Soit l'erreur que ferait le modèle estimé sur un nouvel individu dont le régresseur serait pris au hasard parmi les x_i. Cette erreur est $\frac{1}{n}E[\|y' - X\widehat{\beta}\|^2]$ avec $y' = X\beta^* + u'$ où u' est indépendant de u. Montrer qu'elle vaut $E[\frac{1}{n}RSS] + \frac{2p}{n}\sigma^2$.
   Le terme $\frac{1}{n}RSS$ est l'erreur apparente, qui sous-estime donc l'erreur en prédiction de $2p\sigma^2/n$. C'est pourquoi Mallows a proposé comme mesure de qualité d'un modèle $C_p = \sigma^{-2}RSS + 2p$, σ^2 étant estimé sur le modèle le plus grand.

Exercice 10. Validation croisée. On considère l'estimateur de la variance par validation croisée :

$$\widehat{\sigma}_{CV}^2 = \frac{1}{n}\sum_i (y_i - x_i\widehat{\beta}_{(i)})^2.$$

1. Montrer, en utilisant la formule pour $\widehat{\beta}_{(i)}$, que

$$\widehat{\sigma}_{CV}^2 = \frac{1}{n}\sum_i \frac{\widehat{u}_i^2}{(1-h_i)^2} \tag{II.11}$$

2. Utiliser l'exercice 9 pour calculer son espérance. Comparer $\widehat{\sigma}_{CV}^2$ et $\widehat{\sigma}^2$ dans le cas où les h_i sont égaux.

Exercice 11. Modèle contraint.

1. Soit $\widehat{\beta}$ l'estimateur OLS habituel. Vérifier que l'estimateur OLS sous la contrainte $L\beta = l$, c.-à-d. le minimum de $SS(\beta)$ sous la contrainte $L\beta = l$, est

$$\widehat{\beta}_0 = \widehat{\beta} + (X^TX)^{-1}L^T[L(X^TX)^{-1}L^T]^{-1}(l - L\widehat{\beta}).$$

2. En déduire que $\widehat{y}_0 = X\widehat{\beta}_0$ satisfait :

$$\|\widehat{y} - \widehat{y}_0\|^2 = (L\widehat{\beta} - l)^T \left(L(X^TX)^{-1}L^T\right)^{-1}(L\widehat{\beta} - l).$$

3. Montrer que

$$\|\widehat{y} - \widehat{y}_0\|^2 = \|\widehat{y}_0 - y\|^2 - \|\widehat{y} - y\|^2.$$

   Indication : ne pas utiliser la question précédente.

Exercice 12. Propriété BLUE. Les estimateurs linéaires de β^* sont ceux de la forme

$$\check{\beta} = \Phi(X)y + \Psi(X)$$

où Φ et Ψ sont des fonctions de X. Bien entendu $\widehat{\beta}$ en est un. On a alors la propriété BLUE (Best Linear Unbiased Estimator) :

$\widehat{\beta}$ est de variance minimale dans la classe des estimateurs de β linéaires sans biais.

Démontrer ce résultat de la façon suivante :

1. Montrer que si $\check{\beta}$ est sans biais, nécessairement $\Phi(X)X = I$, et $\Psi(X) = 0$.
2. Exprimer alors $\check{\beta} - \widehat{\beta}$ et $\widehat{\beta} - \beta^*$ en fonction de X et u.
3. En déduire que $Cov(\check{\beta} - \widehat{\beta}, \widehat{\beta}) = 0$, puis que $Var(\widehat{\beta}) \leqslant Var(\check{\beta})$.

Exercice 13. On démontre ici des formules de suppression d'un individu.

1. Prouver le lemme d'inversion matricielle : Soient A, B, C, D quatre matrices, respectivement de taille $n \times n, n \times m, m \times m, m \times n$, alors, si les inverses existent on a

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}.$$

2. Vérifier la formule $X^T X = \sum_i x_i^T x_i$. On peut le faire soit par calcul explicite de chaque coefficient, soit en faisant le produit de matrices $X^T X$ par blocs ; mais le plus simple est sans doute de remarquer que $Id = \sum_i e_i e_i^T$, où les e_i sont les éléments de la base canonique, et que par conséquent $X^T X = X^T(\sum_i e_i e_i^T)X$...
3. En utilisant les deux points précédents, démontrer la formule ($X_{(i)}$ est la matrice déduite de X par suppression de la i-ième ligne)

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i^T x_i (X^T X)^{-1}}{1 - h_i}, \quad h_i = x_i (X^T X)^{-1} x_i^T$$

4. En déduire que :

$$(1 - h_i)^{-1} = 1 + x_i (X_{(i)}^T X_{(i)})^{-1} x_i^T.$$

   En déduire les trois points de la proposition 8.
5. Exprimer $X^T y$ en fonction de $X_{(i)}^T y_{(i)}$ et $x_i^T y_i$, puis démontrer les formules pour les estimées en l'absence du i-ième individu (théorème 9).

Exercice 14. On considère le modèle à deux régresseurs x et z :

$$y = a_1^* x + a_2^* z + u.$$

1. Calculer la matrice de covariance des coefficients estimés et l'exprimer en fonction de $\|x\|$, $\|z\|$ et $\cos(\widehat{x, z})$.
2. Que vaut la corrélation de $\widehat{a}_1$ et $\widehat{a}_2$? Que se passe-t-il si x et z sont orthogonaux ?

Exercice 15. Un goûteur teste des chocolats fabriqués à base de cacao de trois provenances différentes : Côte d'Ivoire, Venezuela, Brésil. Il donne une note pour chaque chocolat qu'il goûte. Les chocolats sont préparés avec des doses de vanilline différentes. Proposer pour cette expérience un modèle de régression avec interaction et un sans interaction. Combien ont-ils de paramètres ? Interpréter leur différence.

Exercice 16. On teste des doses différentes d'engrais dans un champ divisé en parcelles similaires de même taille (une dose par parcelle). On mesure le poids de blé produit à chaque fois et l'on présuppose le gain de production est *proportionnel* à la quantité d'engrais utilisé (la dose).

1. Combien y a-t-il de paramètres à estimer ? Montrer que l'on peut mettre cette expérience sous la forme d'un problème de régression. Combien de colonnes a la matrice X ?
2. On fait maintenant la même expérience mais dans trois champs *différents*. On suppose de plus que l'effet de l'engrais *dépend* du champ. Écrire l'équation de régression. Combien de colonnes a la matrice X ?
3. On suppose que l'engrais a le même effet dans les trois champs, mais que leur productivité en absence d'engrais est toujours différente. Que devient l'équation de régression ? Combien de colonnes a la matrice X ?
4. On a maintenant deux engrais et un seul champ. Combien de colonnes a la matrice X ?

Vérifier que la somme des solutions fait 15.

II.3 Modèles hétéroscédastiques (Moindres carrés généralisés)

II.3.1 Modèle

La différence avec le modèle précédent est que la matrice de covariance des bruits est maintenant différente d'un multiple de l'identité. Il est souvent désigné sous le terme GLS (Generalised Least Squares).

Modèle. *On suppose l'existence d'un vecteur β^*, de $\sigma_* > 0$, de $\Omega_* > 0$ et de variables aléatoires u_i tels que*

$$y = X\beta^* + u,$$
$$E[u] = 0,$$
$$E[uu^T] = \sigma_*^2 \Omega_*.$$

En d'autres termes, pour chaque i :

$$y_i = x_i\beta^* + u_i$$
$$E[u_i] = 0$$
$$Cov(u_i, u_j) = \sigma_*^2 \Omega_{*ij} \qquad \text{(hétéroscédasticité et corrélation des erreurs).}$$

Le paramètre σ_*^2, a priori redondant, est introduit traditionnellement avec l'idée que Ω_* est connu à l'avance et σ_*^2 à estimer, ce qui, on va le voir, nous ramène par une transformation simple au problème précédent. Souvent cependant Ω_* représente directement la matrice de covariance de u (si bien que $\sigma_* = 1$), elle est inconnue, et est paramétrée par un vecteur de taille raisonnable (cf. les deux exemples du § II.3.4).

II.3.2 Réduction au cas $\Omega_* = I$. Estimation de β^* et σ_*^2

Soit R une racine carrée de Ω_*^{-1}, c-à-d $R^T R = \Omega_*^{-1}$; si l'on pose :

$$y' = Ry, \quad X' = RX, \quad u' = Ru$$

on obtient

$$y' = X'\beta^* + u', \quad E[u'u'^T] = \sigma_*^2 R(R^T R)^{-1} R^T = \sigma_*^2 I.$$

On a donc décorrélé et normalisé les observations. On est ramené au problème du § II.2.

Une autre approche qui, on va le voir, conduit aux mêmes conclusions, est de passer par le modèle gaussien $y \sim \mathcal{N}(X\beta^*, \sigma_*^2\Omega_*)$. La matrice Ω_*^{-1} définit une nouvelle métrique sur $\mathbb{R}^n$ qui intervient dans la vraisemblance :

$$\|z\|^2_{\Omega_*^{-1}} = z^T \Omega_*^{-1} z = \sum_{ij} z_i (\Omega_*^{-1})_{ij} z_j.$$

11 - Définition

L'estimateur des moindres carrés généralisés de β^* est l'estimateur du maximum de vraisemblance sous l'hypothèse de normalité de u (c.-à-d. $y \sim \mathcal{N}(X\beta^*, \sigma_*^2\Omega_*)$) :

$$\widehat{\beta}_G = \arg\min_\beta \|y - X\beta\|_{\Omega_*^{-1}} = (X^T\Omega_*^{-1}X)^{-1} X^T \Omega_*^{-1} y.$$

On a bien : $\widehat{\beta}_G(y, X, \Omega_*) = \widehat{\beta}_{OLS}(y', X') = (X'^T X')^{-1} X'^T y'$. En conséquence les résultats du § II.2 s'appliquent :

12 - Proposition

$\widehat{\sigma}^2 = (n-p)^{-1} \|y - \widehat{y}\|^2_{\Omega_*^{-1}}$ est un estimateur sans biais de σ_*^2.

On peut aussi relier ces résultats au § II.2 en remarquant que seule la métrique a changé :

$$\widehat{y}_G = H_G y, \quad H_G = X(X^T\Omega_*^{-1}X)^{-1} X^T \Omega_*^{-1}$$

et H_G est le projecteur orthogonal sur $\mathcal{X}$ pour le produit scalaire $\langle z,t\rangle_{\Omega_*^{-1}} = z^T\Omega_*^{-1}t$.

Attention, la formule de prédiction pour un nouvel individu dont la covariance avec les autres individus est connue et non-nulle diffère de $\widehat{y}_0 = x_0\widehat{\beta}$, du moins si l'on désire prédire le bruit avec, c.-à-d. y_0 et non $E[y_0]$; c'est la formule (II.14) que l'on verra plus loin, qui fait intervenir les corrélations de u_0 avec les u_i (vecteur noté ω).

Donnons l'exemple des **expériences répétées** (ou **données groupées**) : On fait n_i fois la même expérience avec le régresseur x_i ; en notant la réponse moyennée $\bar{y}_i$, on a

$$\bar{y}_i = \beta^* x_i + \bar{u}_i, \qquad Var(\bar{u}_i) = \sigma_*^2/n_i. \tag{II.12}$$

Ω_* est diagonale. L'estimation GLS de β à partir des $\bar{y}_i$ donnera le même résultat que l'estimation OLS sur les données non-moyennées. En revanche l'estimation de σ_* sera moins bonne que si l'on possédait les données non moyennées.

II.3.3 Détection de l'hétéroscédasticité

L'homoscédasicité ne peut être testée en toute généralité car il s'agit de tester tous les coefficients de Ω avec seulement n observations ! On peut imaginer une multitude de tests. La meilleure méthode reste de proposer des modèles hétéroscédastiques plus spécifiques (mixtes, etc.) en fonction de l'idée que l'on se fait des données, et de les tester.

Un test graphique simple est la représentation résidus/valeurs ajustées du § II.2.7 ; une évolution de l'amplitude des $\widehat{u}_i$ quand $\widehat{y}_i$ varie est un indicateur d'hétéroscédasticité.

On peut aussi utiliser le test de Breusch et Pagan [35] qui cherche à détecter si la variance de u_i dépend de x_i. Il fait la régression de $\widehat{\sigma}^{-2}\widehat{u}_i^2$ sur des variables explicatives z_i (en général, $z_i = x_i$) et teste la nullité des coefficients obtenus (procédure `ncv.test` de R) [2]. Si la réponse est non, un estimateur de la matrice de covariance de $\widehat{\beta}$ proposé par Eicker et White est $\widehat{V}_{HCE} = (X^TX)^{-1}\big(\sum_i x_i^T x_i \widehat{u}_i^2\big)(X^TX)^{-1}$, qui peut être amélioré en introduisant les h_i [3].

Si i représente un indice de temps, on utilise parfois le test de Durbin-Watson [39] pour détecter la présence d'une corrélation non-modélisée entre les données ; c'est un test de corrélation entre les résidus dont la statistique est $S = \sum_2^n(\widehat{u}_i - \widehat{u}_{i-1})^2 / \sum_1^n \widehat{u}_i^2$. Si S est trop grande (resp. petite) il y a alors une corrélation significativement négative (positive) entre les résidus.

II.3.4 Estimation de Ω_*

Il est totalement désespéré d'estimer Ω_* sans contrainte supplémentaire car cette matrice contient plus de coefficients que de données dont on dispose. On postulera donc toujours pour Ω_* une forme très particulière avec peu de paramètres (cf. (II.12), les exemples de ce paragraphe et le § II.3.5).

Une méthode générale. Si l'on dispose d'un estimateur $\Omega(\beta, y, X)$ de Ω_* en fonction des données et de β^*, β^* et Ω_* peuvent être alors estimés en répétant les deux opérations suivantes :

- ► pour une valeur estimée $\widehat{\Omega}$ de Ω_* calculer : $\widehat{\beta} = (X^T\widehat{\Omega}^{-1}X)^{-1}X^T\widehat{\Omega}^{-1}y$
- ► puis ensuite estimer Ω_* à l'aide de $\widehat{\beta}$: $\widehat{\Omega} = \Omega(\widehat{\beta}, y, X)$.

La convergence de la méthode dépend de chaque situation particulière.

Exemple : « Seemingly unrelated regression ». Soit le modèle

$$\begin{pmatrix} y \\ y' \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & X' \end{pmatrix}\begin{pmatrix} \beta \\ \beta' \end{pmatrix} + \begin{pmatrix} u \\ u' \end{pmatrix}, \quad Var\begin{pmatrix} u \\ u' \end{pmatrix} = \begin{pmatrix} v_{11}Id & v_{12}Id \\ v_{12}Id & v_{22}Id \end{pmatrix}.$$

2. Dans un même esprit le test de White [79] compare $S = n^{-1}\sum_i(\widehat{\sigma}^2 - \widehat{u}_i^2)(x_i^T x_i - n^{-1}X^TX)$ à 0. C'est un test de corrélation entre les carrés des résidus et les régresseurs. Concrètement, la matrice S vectorisée s'écrit $n^{-1}\sum v_i$, où $v_i \in \mathbb{R}^{p(p+1)/2}$ contient donc les termes $(\widehat{\sigma}^2 - u_i^2)(x_{ij}x_{ik} - n^{-1}(X^TX)_{jk})$, j et k variant. Le test compare $(\sum v_i)^T(\sum v_i v_i^T)^{-1}(\sum v_i)$ à un $\chi^2_{p(p+1)/2}$.

3. Pour des références, voir la page Wikipedia : *Heteroscedasticity-consistent standard errors*.

Par exemple y_i et y'_i sont deux mesures différentes, ou différées, de l'activité commerciale du pays i. Une estimée de β et β' permet alors d'en déduire une des v_{ij} à partir $\widehat{u}$ et $\widehat{u}'$, ce qui conduit donc à un $\widehat{\Omega}$.

Exemple : Variance dépendant d'une modalité et proportionnelle à une variable. On ensemence des pots avec les mêmes graines mais des terreaux différents, variable t, et en éclairant constamment chaque plante avec une intensité lumineuse a différente pour chacune ; on mesure la hauteur y de la plante au bout d'un mois

$$y_i = \alpha_{t_i} + \beta a_i + u_i, \quad Var(u_i)^{1/2} = \sigma(\gamma_{t_i} + a_i^{\delta_{t_i}}).$$

La variance dépend donc du type de terreau et du niveau d'éclairage. La commande R correspondante sera (bibliothèque `nlme`) :

```
gls(y∼t+a,weights=varConstPower(form=∼a|t))
```

Si la variance ne dépend que de t, faire `weights=varIdent(form=∼1|t)`.

Exemple : Données longitudinales. On s'intéresse à savoir si le labour a une influence sur la présence de carbone dans le sol [4]. On prélève des carottes dans divers champs et la mesure y_{ij} est le taux de carbone à la profondeur t_j de la i-ième carotte. Les variables explicatives sont le taux d'humidité et l'année. En raison de la corrélation présente le long de la carotte on postule le modèle suivant (en fait il y a deux modèles : un avec labour et un sans labour)

$$y_{ij} = x_i\beta + \sum_{k=1}^{K} \gamma_k t_j^k + v_{ij} + u_{ij}$$
$$E[uu^T] = \sigma_u^2 Id, \quad E[v_{ij}v_{ik}] = \sigma_v^2 \exp\{-\alpha|t_j - t_k|\}, \quad E[v_{ij}v_{i'k}] = 0, \quad i' \neq i.$$

Le terme polynomial (p.ex. $K = 1$) explique une tendance régulière de variation du taux de carbone en fonction de la profondeur ; le terme v_{ij}, ajouté à u_{ij}, exprime une corrélation additionnelle entre prélèvements proches pour le même individu.

Chaque estimée $\widehat{\beta}, \widehat{\gamma}$ de β^*, γ^* conduit à une estimée des $u_{ij} + v_{ij}$, qui fournissent à leur tour une estimée de σ_u, σ_v et α. On a donc ainsi fabriqué une fonction $\widehat{\Omega}(\beta, y, X)$.

Filtrage par krigeage. Soit le modèle de régression habituel où les y_i sont typiquement des mesures prises chacune en un point ξ_i du plan, c.-à-d. un champ spatial (intensité lumineuse en un point d'une image, mesure de pollution en un endroit, etc.) [6]. On se propose d'exprimer la corrélation comme une fonction paramétrée des localisations, par exemple

$$y_i = x_i\beta + u_i, \qquad i = 1, \dots n,$$
$$E[u_iu_j] = c\exp(-b\|\xi_i - \xi_j\|^a) + \sigma_0^2\delta_{ij}. \tag{II.13}$$

Nous sommes encore dans le cadre précédent ; les quatre paramètres a, b, c, σ_0^2 devront être estimés. Le deuxième terme de (II.13) peut paraître étrange puisque l'on a $E[(u_i - u_j)^2] = O(\|\xi_i - \xi_j\|^a) + 2\sigma_0^2$, qui ne tend pas vers 0 quand $\xi_i \to \xi_j$; on peut considérer que le premier terme représente le champ spatial lui-même, tandis que $\sigma_0^2\delta_{ij}$ représente une erreur supplémentaire due à la mesure.

Souvent dans les applications seul le régresseur constant est considéré mais ce n'est pas toujours le cas. Le but du krigeage est de prédire la réponse y_0 en un nouveau point ξ_0 en prenant en compte les variables explicatives x_0 et en exploitant les corrélations existant avec les y_i. Ici $\sigma_* = 1$ et on note

$$y = (y_i)_{1\leqslant i\leqslant n}, \quad u = (u_i)_{1\leqslant i\leqslant n}, \quad \Omega = E[uu^T], \quad \omega = E[uu_0].$$

On suppose dans la suite que le régresseur constant est pris en compte dans x. Si β^* et Ω étaient connus, l'estimateur naturel de y_0 serait, sous l'hypothèse gaussienne [5], son espérance sachant les y_i soit

4. [30]. Nous simplifions ici beaucoup : En réalité F.J. Breidt utilise des fonctions splines et les termes correspondant aux γ_k dépendent également des variables explicatives, le tout dans un cadre de modèles mixtes.

5. Si (X, Y) est un vecteur gaussien centré dans $\mathbb{R}^{n+p}$, on a $E[Y|X] = R_{YX}R_{XX}^{-1}X$.

$\widehat{y_0^*} = E[x_0\beta^* + u_0|u] = x_0\beta^* + \omega^T\Omega^{-1}u$. Comme β^* et Ω sont inconnus on choisit de les estimer et il vient

$$\widehat{y_0} = x_0\widehat{\beta} + \omega^T\widehat{\Omega}^{-1}(y - X\widehat{\beta}). \tag{II.14}$$

Il ne reste donc plus qu'à trouver $\widehat{\beta}$ et $\widehat{\Omega}$, c.-à-d. dans l'exemple (II.13) estimer β, a, b, c et σ_0^2. En pratique on estime d'abord Ω puis on utilise $\widehat{\beta}_G$.

Parenthèse. Rappelons que *le régresseur constant est pris en compte.* Il est d'usage, dans la littérature de krigeage, de remarquer que $\widehat{y_0} = \sum p_i y_i$ où le vecteur p est solution d'une des deux équations suivantes en (p, q) au choix [6]

$$\begin{pmatrix} \Omega & X \\ X^T & 0 \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} \omega \\ x_0^T \end{pmatrix} \quad \text{ou} \quad \begin{pmatrix} \Gamma & X \\ X^T & 0 \end{pmatrix}\begin{pmatrix} p \\ q' \end{pmatrix} = \begin{pmatrix} \gamma \\ x_0^T \end{pmatrix} \tag{II.15}$$

$\Gamma_{ij} = E[(u_i - u_j)^2]$, $\gamma_i = E[(u_i - u_0)^2]$.

La matrice Γ est appelé le variogramme. Dans le domaine des processus spatiaux, il apparaît souvent plus naturel de travailler avec cette matrice plutôt qu'avec Ω pour des raisons de modélisation. Quelques remarques :
1/ $\sum_i p_i = 1$ puisque $X^T p = x_0^T$ (cette propriété est perdue si le régresseur constant n'est pas pris en compte, et la deuxième équation matricielle également). Les p_i ne sont pas nécessairement $\geqslant 0$.
2/ Il est immédiat de vérifier que (II.15) correspond à la solution du problème en p

$$\min Var(y_0 - \sum_i p_i y_i) \quad \text{sous} \quad \sum p_i x_i = x_0$$

sachant que $Cov(y_i, y_j) = \Omega_{ij}$ et $Cov(y_0, y_i) = \omega_i$, $i, j \geqslant 1$.
3/ p est également solution de (II.15) modifié en remplaçant u par y dans les définitions de Ω, ω, Γ et γ.

L'estimation de Ω (ou Γ) se fait généralement par le biais d'un modèle du type $\Omega_{ij} = \varphi(\theta_i, \theta_j)$ où θ_i est un vecteur de variables explicatives ; en dehors de (II.13) voici un autre exemple [70]

$$\Omega_{ij} = c\exp\big(-\textstyle\sum_k c_k|x_{ik} - x_{jk}|^\alpha\big) + \sigma^2\delta_{ij}$$

ou encore $\Omega_{ij} = \rho(\|\xi_i - \xi_j\|)$ où la fonction ρ est estimée par un estimateur non paramétrique p.ex.

$$\widehat{\rho}(h) = \frac{1}{N(h,\delta)} \sum_{h-\delta<|\xi_i-\xi_j|<h+\delta} \widehat{u}_i\widehat{u}_j$$

et $N(h, \delta)$ est le nombre de termes dans la somme, δ un paramètre à choisir. Le problème est d'obtenir une matrice positive à la fin.

Ou le modèle « sphérique » $\Omega_{ij} = \alpha g(\|\xi_i - \xi_j\|/\beta)$

$$g(x) = 1 + 1_{x<1}(-\tfrac{3}{2}x + \tfrac{1}{2}x^3)$$

(la fonction g est à dérivée continue) ou le modèle exponentiel $g(x) = e^{-x}$. On trouvera des compléments dans [20].

II.3.5 Modèles mixtes

C'est le modèle de régression

$$y = X\beta + Z\gamma + u, \quad \gamma \sim \mathcal{N}(0, G), \quad u \sim \mathcal{N}(0, \sigma^2 Id) \tag{II.16}$$

où X et Z sont des matrices connues (régresseurs), β est le paramètre et γ est un bruit vectoriel indépendant de u. G est typiquement une matrice diagonale. On peut remplacer $\sigma^2 Id$ par une matrice plus

6. Utiliser la propriété suivante : Soit A, B, C, D quatre matrices de dimensions adéquates, si les inverses existent

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E & -A^{-1}BF \\ -FCA^{-1} & F \end{pmatrix}, \quad F = (D - CA^{-1}B)^{-1}, \quad E = A^{-1} + A^{-1}BFCA^{-1}.$$

générale. Noter que Z a un nombre a priori faible de colonnes, et donc le bruit $(Z\gamma)_i$ engendré par γ est très corrélé d'une donnée à l'autre, contrairement à u_i. On a

$$y \sim \mathcal{N}(X\beta, V), \quad V = \sigma^2 Id + ZGZ^T. \tag{II.17}$$

On peut voir à l'inverse cette modélisation comme $y \sim \mathcal{N}(X\beta + Z\gamma, \sigma^2 Id)$ avec l'introduction d'une information Bayésienne sur une partie des coefficients (les γ_j).

Un point de vue plus pragmatique est d'y voir une possibilité d'estimer d'un modèle linéaire quand la matrice de régression, ici $(X|Z)$, a beaucoup de colonnes, voire plus de colonnes que de lignes. Même si G est inconnu, il pourra en pratique être estimé via une modélisation paramétrique dont la plus simple est $G = \sigma_\gamma^2 Id$.

En résumé : *Le modèle mixte est une formulation particulière de modèle hétéroscédastique ; elle permet en particulier de proposer un modèle de complexité intermédiaire entre le modèle complet $y \sim \mathcal{N}((X|Z)\beta, \sigma I)$ (qui a trop de paramètres) et le modèle $y \sim \mathcal{N}(X\beta, \sigma I)$ qui est trop simple.*

Exemple : Données groupées (random block effects). Supposons que l'on a rassemblé p groupes de données obtenues dans des conditions différentes. Par exemple chaque groupe peut représenter une série d'expériences (test de cocktails, traitements médicaux, etc.) faites sur un sujet (différent d'un groupe à l'autre). À l'intérieur de chaque groupe on ne peut pas considérer les mesures comme indépendantes car elles ont en commun des conditions expérimentales spécifiques (le goûteur, le cobaye, etc.). En désignant par x_e, $e = 1, \dots n_e$, les régresseurs de l'expérience numéro e (composition du cocktail, teneur en sucre, etc.) et s, $s = 1, \dots n_s$ le sujet (goûteur), on pourra choisir le modèle :

$$y_{se} = x_e\beta + \gamma_s + u_{se}, \quad \gamma_s = \mathcal{N}(0, \sigma_g^2).$$

On voit que l'effet aléatoire se traduit ici par un *biais* variant aléatoirement d'un sujet à l'autre (certains goûteurs sont plus sévères, etc.) ; en reprenant les notations précédentes, et en supposant que chaque sujet fait toutes les expériences, Z est $(n_e n_s) \times n_s$:

$$Z = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{1} \end{pmatrix}, \quad \gamma = \mathcal{N}(0, \sigma_g^2 I_{n_s})$$

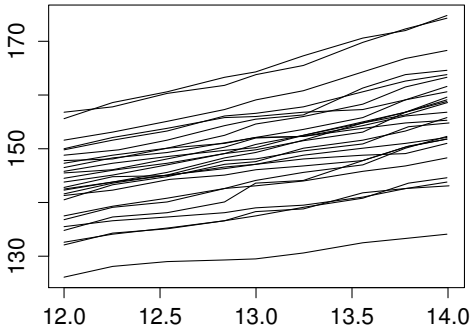
où $\mathbf{0}$ et $\mathbf{1}$ sont des vecteurs de 0 et de 1. Ici, σ_g mesure la variabilité de la réponse due au changement de conditions expérimentales (fluctuations de la sévérité d'un goûteur à l'autre). Dans le cas des cocktails, il est clair le modèle complet n'est pas beaucoup plus intéressant d'un point de vue pratique que le modèle mixte, puisqu'on ne s'intéresse pas aux sujets individuellement.

Noter que dans cet exemple $(X|Z)$ *est de rang déficient*, à cause de la colonne de 1 dans X ; en effet, en raison du caractère aléatoire des γ_i dans la modélisation, il est important de préserver la symétrie : l'effet du retrait d'une colonne redondante dépendrait ici de la colonne choisie. Si l'on cherche à estimer les γ_i dans le modèle $y \sim \mathcal{N}((X|Z)\binom{\beta}{\gamma}, \sigma I)$, il est naturel d'ajouter la condition $\sum \gamma_i = 0$, liée au fait que les γ_i du modèle mixte sont centrés, et la solution redevient unique.

Exemple : Fluctuations sur les paramètres d'un modèle longitudinal [66]. Les auteurs cherchent à proposer un modèle de croissance pour des enfants de 12 à 14 ans. Il y a 26 enfants. On mesure la taille y_{ij} de l'enfant i à l'âge t_{ij}, $j = 1, \dots J = 9$ (les mesures sont prises tous les trois mois). Les auteurs postulent le modèle polynomial (le choix des ordres 4 et 2 est de nature expérimentale)

$$y_{ij} = \sum_{k=0}^{4} \beta_k t_{ij}^k + \sum_{k=0}^{2} \gamma_{ik} t_{ij}^k + u_{ij}, \qquad (\gamma_{.0}, \gamma_{.1}, \gamma_{.2}) \sim \mathcal{N}(0, G). \tag{II.18}$$

Si J est petit, il est hors de question d'estimer un polynôme d'ordre 4 par enfant, et cela présente peu d'intérêt car l'interprétation du lot de paramètres obtenus exigera une nouvelle analyse statistique. Les auteurs choisissent donc le modèle (II.18). Les β_k représentent le polynôme moyen tandis que les γ_{ik} servent à modéliser la variabilité d'un individu à l'autre. L'estimateur de G quantifie cette variabilité.



L'analyse avec la commande

`mod=lme(taille~poly(t,4),random=~1+t+I(t^2)|sujet)`,

(bibliothèque `nlme`) donne les estimées (effets, fixes, variances, corrélations, avec la convention $G_{ij} = \sigma_i\sigma_j r_{ij}$ pour $i \neq j$)

β_0	β_1	β_2	β_3	β_4	σ_0	σ_1	σ_2	σ_u	r_{01}	r_{02}	r_{12}
149	6,2	1,1	0,47	−0,34	8	1,7	0,8	0,47	0,61	0,22	0,66

Notons les fortes corrélations, et la cohérence de $r_{01} > 0$ avec la figure. Les intervalles de confiance pour ces quantités s'obtiennent avec `intervals(mod)`. L'estimation du modèle $y_{ij} = \sum_{k=0}^{4} \beta_k t_{ij}^k + u_{ij}$, donne bien les mêmes estimées de la partie fixe.

On peut voir ce modèle comme un **modèle hiérarchique** car le coefficient de t^k pour l'enfant i est $\beta_k + \gamma_{ik}$, ce qui revient à postuler un modèle de régression pour ce coefficient.

Pour additionner plusieurs effets aléatoires indépendants il faut faire une liste, par exemple

`mod1=lme(taille~poly(t,4),random=list((~1+t|sujet),(~I(t^2)-1|sujet)))`

revient à imposer $r_{02} = r_{12} = 0$.

On peut également utiliser la bibliothèque `lme4` qui donne de bons résultats en estimation ; la commande sera `mod=lmer(taille~poly(t,4)+((1+t+I(t^2))|sujet))`. Pour un modèle avec uniquement des effets aléatoires du type $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + u_{ijk}$ où les trois effets sont indépendants (les paramètres sont $(\mu, \sigma_u, \sigma_\alpha, \sigma_\beta, \sigma_\gamma)$) on fera `lmer(y~1+(1|A)+(1|B)+(1|A:B))`. Ceci est beaucoup plus difficile à réaliser avec `lme` qui est adapté aux effets emboîtés, cf. § III.4.6 et § III.4.7.

Pour les cocktails, on fera `lmer(y~x+(1|sujet))` ou `lme(y~x,random=~1|sujet)`.

Tester un facteur emboîtant. Supposons que les enfants se séparent en deux groupes, «végétarien» et «non-végétarien» (variable v) et que l'on veuille tester l'importance du régime. Supposons ici pour simplifier que l'on considère une croissance linéaire en temps. Une solution pour tester v est de considérer le modèle avec une droite par enfant `lm(taille~t*sujet)`

$$y_{ij} = \alpha_i + \beta_i t_{ij} + u_{ij}$$

et tester $\frac{1}{n_V}\sum_{i\in V}\beta_i = \frac{1}{n-n_V}\sum_{i\notin V}\beta_i$, où n_V est le nombre de végétariens et V est l'ensemble des végétariens. Noter que comme la variable «sujet» détermine v (emboîtement), ajouter v à ce modèle ne le modifie pas ; on ne peut donc pas faire un test basé simplement sur l'ajout du facteur au modèle ; en revanche ceci peut être fait dans un cadre de modèle à effets aléatoires en testant

`modV=lme(taille~(poly(t,4)*v,random=~1+t+I(t^2)|sujet)`

(on a repris la dépendance polynomiale en temps) contre `mod` défini plus haut (nous verrons les tests au § III.4.7). Noter que comme le terme à effets fixes peut être vu comme une valeur moyenne sur les sujets, tester sa dépendance en v est conceptuellement similaire au premier test présenté.

Bilan. On vient de voir que le modèle mixte permet d'obtenir des résultats d'analyse plus synthétiques

qu'un modèle complet.

Il permet de tester facilement un facteur emboîtant.

Il permet aussi de juger de l'importance des régresseurs Z dans des situations où l'on ne peut pas estimer le modèle complet : si le modèle mixte est significativement meilleur que le modèle à effets fixes simple, alors les régresseurs Z jouent un rôle significatif. On verra au § III.4.7 qu'une application typique est de tester les interactions compliquées en analyse de la variance.

Estimation. Les paramètres sont (β, σ, G), où G a une forme paramétrique, typiquement diagonale. L'estimation est faite soit par maximum de vraisemblance sur la base de l'équation (II.17), soit par la méthode REML qui consiste en un premier temps à projeter y sur l'orthogonal de X, $y^* = Ky$, à utiliser ces nouvelles observations dont la loi ne dépend plus de β pour estimer (σ, G), et en un second temps à estimer β classiquement (GLS), voir [18] § 6.6 et [1] chap.11 § 1.6. Le REML est souvent préféré car comme y^* est de dimension effective $n - p$ (on le représente en fait dans une base de $\mathcal{X}^\perp$), les variances estimées seront naturellement mieux normalisées.

II.3.6 Exercices

Exercice 1. On considère le modèle de régression

$$y_i = ax_i + u_i, \quad i = 1, \dots N$$

avec : $E[u_i] = 0$, $Var(u_i) = \sigma_i^2$, $\mathrm{Cov}(u_i, u_j) = 0$, $i \neq j$. x_i et a sont scalaires. Donner l'expression des estimateur OLS et GLS de a et comparer leur variance.

Exercice 2. On recueille J séries de mesures de modèle

$$y_{ij} = \mu + u_{ij}, \quad \mathrm{Var}(u_{ij}) = \sigma_j^2, \quad i = 1, \dots n, \quad j = 1, \dots J.$$

Les bruits sont donc décorrélés mais de variance différente connue. Mettre sous forme homoscédastique par un changement de variable adéquat puis en déduire l'expression de l'estimateur de μ.

Les variances sont inconnues. Écrire la commande R faisant l'estimation.

Exercice 3. (D'après [57]) Des vaches donnent naissance à des veaux, issus de 4 taureaux. Les vaches proviennent de deux troupeaux. Chaque expérience est un accouchement. Les régresseurs sont l'âge de la vache, le sexe du veau, le taureau (variable catégorielle), et le troupeau. La réponse est la difficulté que la vache a eue à vêler (note donnée par un technicien). Le but principal de l'étude est de comprendre la variabilité du résultat d'un taureau à l'autre. Il y a 28 individus.

1. Calculer le nombre de paramètres du modèle additif utilisant toutes les variables.
2. Proposer un modèle à effets aléatoires et donner son nombre de paramètres. Justifier le choix de ce modèle en termes d'interprétations de la régression et de son utilisation.
3. Ecrire les commandes `lme()` et `nlme()` correspondant à cette analyse.
4. Il y a deux races de taureau. On s'intéresse également à la variabilité de l'effet « taureau » pour chaque race séparément. Ecrire le modèle correspondant.

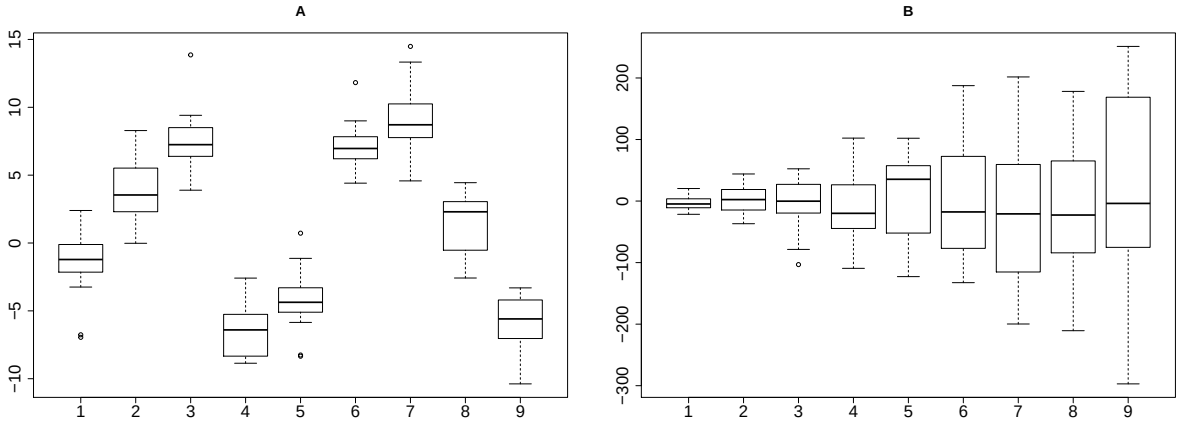
   La commande est `lme(y~Tr+A+S,random=list(Ta=pdDiag(~0+Race)))`, qui indique que Ta est un effet aléatoire avec une matrice de covariance diagonale dont les coefficients ne dépendent que de Race. Par exemple, la syntaxe `lme(y~0+Tr,random=(~1|Ta))` équivaut à `lme(y~0+Tr,random=list(Ta=pdIdent(~1)))`.

Exercice 4. On reprend l'exemple des goûteurs. On suppose que l'on dispose en outre d'une variable « Age du goûteur » à deux modalités.

1. Comment modifier le modèle pour voir si les jeunes sont plus sévères que les vieux ?
2. Comment modifier le modèle pour voir s'il y a une plus grande disparité de sévérité chez les jeunes que chez les plus âgés ?
3. Ecrire la commande `lme()` pour réaliser l'analyse.

Exercice 5. Soient les deux ensembles de données, correspondant aux réponses de 20 sujets (variable j) à 9 stimuli (variable s). Il s'agit de deux expériences n'ayant rien à voir.

Ces données sont ici représentés graphiquement, la réponse y_{sj} étant en ordonnée et la variable explicative s en abscisse :



Soient les modèles candidats :

(1) $y_{sj} = \beta_j + u_{sj}, \quad u_{sj} \sim \mathcal{N}(0, \sigma_1^2)$

(2) $y_{sj} = \alpha + u_{sj}, \quad u_{sj} \sim \mathcal{N}(0, s^2\sigma_1^2)$

(3) $y_{sj} = \alpha + \beta_s + u_{sj}, \quad \beta_s \sim \mathcal{N}(0, \sigma_0^2), \quad u_{sj} \sim \mathcal{N}(0, \sigma_1^2)$

(4) $y_{sj} = \alpha + \beta_j + u_{sj}, \quad \beta_j \sim \mathcal{N}(0, \sigma_0^2), \quad u_{sj} \sim \mathcal{N}(0, \sigma_1^2)$

(5) $y_{sj} = \beta_{sj} + u_{sj}, \quad u_{sj} \sim \mathcal{N}(0, \sigma_1^2).$

1. Au vu des figures seulement, attribuer à (A) et (B) le modèle qui convient à chacun, choisi parmi les cinq.
2. Donner pour ces deux modèles une estimation visuelle (très) approximative des paramètres.

II.4 Moindres carrés totaux (Errors in variables, total least squares)

Modèle. Il a pour but de prendre en compte du bruit sur les régresseurs :

$$y_i = x_i\beta^* + u_i \tag{II.19}$$

$$z_i = x_i + v_i \tag{II.20}$$

où l'on observe les (y_i, z_i) mais pas x_i, qui est qualifié de «variable latente». Les u_i et v_i sont des bruits indépendants de variance σ_u^2 et Σ_v^2. Par exemple dans (I.1), la mesure de $\log p_i$ est sans doute autant entachée d'erreur que celle de la température. L'estimateur OLS de β^* et X s'obtient par minimisation en X et β de

$$\sum_i \sigma_u^{-2}(y_i - x_i\beta)^2 + (z_i - x_i)\Sigma_v^{-2}(z_i - x_i)^T. \tag{II.21}$$

Si $\sigma_u = 1, \Sigma_v = Id$, c'est la somme des carrés des distances des points d'observation (z_i, y_i) aux points de la droite (de l'hyperplan) de régression $(x_i, x_i\beta)$, et le minimum sur les x_i est donc la somme des carrés des distances à la droite. On peut être alors tenté de maximiser la vraisemblance de $(y_i, z_i)_{i\leqslant n}$ par rapport aux paramètres $(\beta, \sigma_u, \Sigma_v, (x_i)_{i\leqslant n})$; on trouve alors $\Sigma_v = 0$ avec $X = Z$, une vraisemblance infinie et $\widehat{\beta} = \widehat{\beta}_{OLS}$ [75]. Même si Σ_v est connu, le maximum de vraisemblance pose des problèmes de consistance ([8] p.104), ceci est dû au nombre déraisonnable de paramètres (les x_i).

Une option statistiquement plus correcte est de considérer que les x_i sont i.i.d. $\mathcal{N}(m, R)$ et d'estimer

$(\beta, \sigma_u, \Sigma_v, m, R)$; le même problème d'identifiabilité apparaît cependant, et l'on doit supposer Σ_v connu, cf. [8] § 2.2.1. La solution du maximum de vraisemblance pour (β, σ_u, m, R) est alors celle de l'estimateur des moments déduit de $E[zy] = E[z^T x\beta^*] = (E[z^T z] + \Sigma_v^2)\beta^*$:

$$\widehat{\beta} = (Z^T Z + \Sigma_v^2)^{-1} Z^T y.$$

Σ_v doit donc être estimé par ailleurs, sur d'autres données, ou grâce à des répétitions de z_i correspondant au même individu x_i (Si u_i ou v_i n'est pas gaussien, alors l'identifiabilité revient [69]).

Une autre voie de sortie apparaît si l'on dispose d'une variable w_i, dite instrumentale, de même dimension que x_i, et indépendante de u_i et v_i, car alors (II.19, II.20) implique $E[w_i^T y_i] = E[w_i^T x_i]\beta^* = E[w_i^T z_i]\beta^*$, ce qui permet d'estimer β^* par $\widehat{\beta} = (W^T Z)^{-1} W^T y$.

On trouvera des compléments dans [8] et [19].

Soulignons que la prise en considération du bruit sur les observations est un problème très général : si, par exemple, l'on observe un AR(1) bruité, $x_n = ax_{n-1} + u_n$, $y_n = x_n + v_n$, l'estimation autorégressive sur y_n conduira à un AR d'ordre arbitrairement grand : on ne peut pas oublier v_n. La solution est ici d'estimer un ARMA(1,1) qui est bien la nature de y.

II.5 Régression non-paramétrique et moindres carrés

Nous ne traitons pas ici de ce problème dans toute sa généralité ; on ne fera pas non plus une présentation rigoureuse de la théorie ; voir p.ex. [10].

Le modèle est le suivant où le paramètre à estimer est la fonction f inconnue :

$$y_i = f(\xi_i) + u_i, \quad u_i = \mathcal{N}(0, \sigma^2), \quad \xi_i \in [0, 1]. \tag{II.22}$$

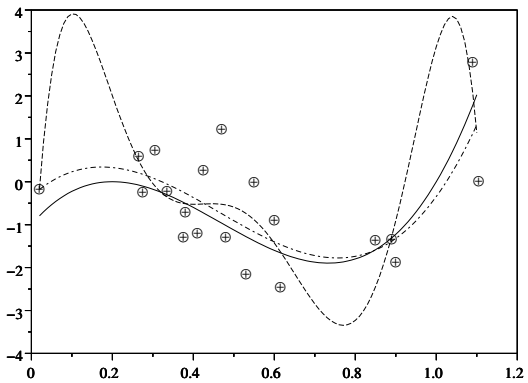
Il s'agit d'un problème non paramétrique car l'ensemble des fonctions candidates n'est pas un espace de dimension finie.

II.5.1 Première approche : la régression polynômiale

Faisons apparaître sur un exemple simulé simple les problèmes rencontrés. On dispose des 20 paires (ξ_i, y_i) représentées sur la figure par des cibles. Elles suivent le modèle (II.22) sauf que l'intervalle de variation de ξ est plus grand. La fonction f est un polynôme d'ordre 3, $f(x) = (5x-1)^2(x-1)$, représenté en trait plein ; on a pris $\sigma = 1$. Si l'ordre est effectivement connu, l'estimation peut se faire par une régression habituelle avec le modèle

$$y_i = \beta_1 + \beta_2 \xi_i + \beta_3 \xi_i^2 + \beta_4 \xi_i^3 + u_i.$$

En général l'ordre n'est pas connu (et même l'hypothèse « f polynômiale » n'est qu'une approximation) et la question de l'ordre à utiliser se pose. La figure montre les estimées pour des ordres 3 et 6 (courbes en pointillés).



On voit que l'estimée avec l'ordre 6 est très mauvaise ; ceci vient du fait que les coefficients supplémentaires ont été utilisés pour approcher davantage les données (bruitées) ce qui a induit un écart important à

la vérité aux endroits où les observations se font rares ; si l'on augmente l'ordre, le polynôme estimé va s'approcher de plus en plus des points d'observation en ayant un comportement très chaotique entre ces derniers, c'est ce que l'on appelle le *surajustement (overfitting)*, il est dû à la *surparamétrisation*. La difficulté est donc de trouver un ordre (taille du modèle) raisonnable.

Estimation du degré par validation croisée. L'idée est d'essayer de choisir la valeur du degré d qui minimisera l'erreur de prédiction. Pour estimer cette erreur, le plus simple est d'utiliser l'estimateur CV (également appelé PRESS : Predicted Residual Sum of Squares) : pour tout i, calculer le modèle $\widehat{\beta}_{(i)}$ (cf. § II.2.5) puis

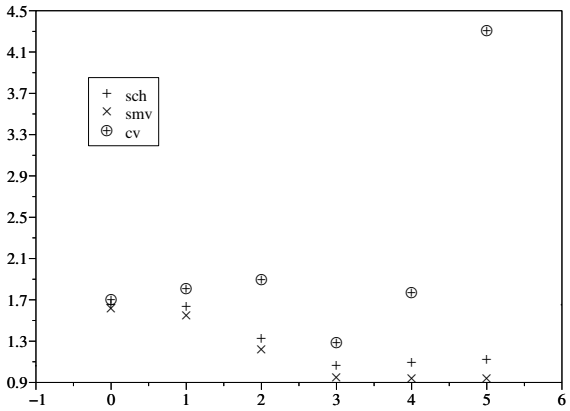
$$CV(p) = n^{-1} \sum_i (y_i - x_i \widehat{\beta}_{(i)})^2$$

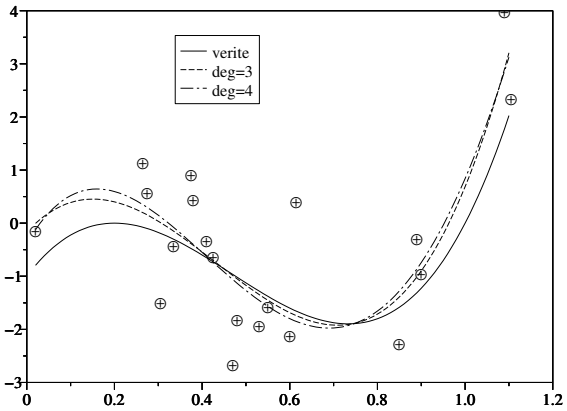
où la dépendance en d est implicite via la dimension du régresseur, $p = d + 1$. Il est essentiel d'ôter l'individu i à chaque fois car sinon on aurait une fonction décroissante de p et il serait finalement choisi trop grand (typiquement égal à n). En utilisant l'exercice 10 p. 22, le critère à minimiser devient

$$\boxed{CV(p) = n^{-1} \sum_i \frac{\widehat{u}_i^2}{(1 - h_i)^2}}$$

Les deux figures suivantes illustrent la méthode. La première montre l'évolution de $\sqrt{CV(p)}$ en fonction de d ainsi que celle de $\widehat{\sigma}$ et de l'estimateur au maximum de vraisemblance $\widehat{\sigma}_{MV} = \sqrt{RSS/n}$. $\widehat{\sigma}_{MV}$ est une fonction décroissante de d car c'est la norme de la projection sur des espaces emboîtés. $\widehat{\sigma}$ n'est visiblement pas non plus une mesure très satisfaisante. En pratique $CV(p)$ avoisine son minimum sur un plateau de largeur réduite où les estimées diffèrent assez peu.

La seconde figure montre les polynômes estimés pour $d = 3$ et $d = 4$, qui semblent être les deux seules valeurs acceptables.





Il a été remarqué que le critère CV n'est pas invariant par rotation au sens où si Q est une matrice de rotation, le calcul de CV sur les données (Qy, QX) (qui satisfont le modèle avec le même β) ne donne pas le même résultat (noter que ce défaut d'invariance n'est pas forcément un défaut). La solution proposée est de prendre le Q qui rend les h_i égaux, ces derniers valent alors p/n (car la somme de h_i reste inchangée) et l'on obtient le critère de validation croisée généralisée [44]

$$\boxed{GCV(p) = \frac{n}{(n-p)^2} RSS}$$

qui est beaucoup plus simple à calculer. Le paragraphe suivant décrit une approche classique qui utilise la base de Fourier plutôt que les polynômes. D'autres choix sont encore possibles comme on le verra plus loin.

II.5.2 Approche par estimation des coefficients de Fourier

Revenons au modèle (II.22). Le paramètre est ici la fonction f ; on est donc en dimension infinie. Une façon de le visualiser plus clairement est de passer par la transformée de Fourier de f

$$\beta_j = \int_0^1 e_j(\xi) f(\xi)\, d\xi, \qquad f(\xi) = \sum_{j \geqslant 0} \beta_j e_j(\xi) \tag{II.23}$$

$$e_0(\xi) = 1, \quad e_{2j}(\xi) = \sqrt{2} \cos 2\pi j \xi, \quad e_{2j+1}(\xi) = \sqrt{2} \sin 2\pi (j+1)\xi, \quad j = 1, 2, \ldots \tag{II.24}$$

On a alors

$$y = X\beta + u, \qquad X_{ij} = e_j(\xi_i) \tag{II.25}$$

qui est la forme habituelle (vu l'absence d'ambiguité, on a supprimé dans ce paragraphe l'étoile qui désignait précédemment le vrai paramètre par opposition au paramètre générique). Noter que la matrice X^T ne peut être de rang plein car elle a n colonnes et une infinité de lignes ; toute estimée OLS sera une fonction qui vaut y_i en ξ_i, donnant ainsi un résidu nul, ce qu'on pouvait deviner tout de suite au vu de (II.22). Cette estimée sera toujours mauvaise, sauf si $\sigma = 0$, car la fonction obtenue sera très irrégulière. Le but sera ici de construire une estimée qui sera bonne si f est régulière (*la situation étant désespérée sans hypothèse supplémentaire sur* f).

On va construire un estimateur biaisé, mais dont les performances seront très supérieures à OLS dans le cas où beaucoup de β_j sont petits (ce qui correspond à f régulière) et très légèrement dégradées sinon. Rappelons que des intégrations par parties dans (II.23) montrent que si f admet q dérivées intégrables, $|\beta_j| < Cj^{-q}$.

Méthode de projection. On se restreint aux estimateurs (biaisés) satisfaisant :

$$\widehat{\beta}_j = 0, \qquad |j| > j_0$$

pour un certain j_0 inférieur à n. Une fois j_0 choisi, le problème est alors un problème de régression purement paramétrique puisqu'il reste à estimer par moindres carrés les β_j pour $|j| \leqslant j_0$.

Ceci correspond à l'*introduction d'une hypothèse supplémentaire* : la suite des β_j tend « rapidement » vers 0. C'est une hypothèse de régularité de f.

L'estimation de j_0 par validation croisée se fait comme précédemment.

La validité théorique de l'approche par validation croisée généralisée a été démontrée par Polyak et Tsybakov [67].

II.5.3 Aspects pratiques

Au problème du choix du nombre de fonctions de base (c.-à-d. du degré, ou de j_0) s'ajoute celui de la base elle-même. En particulier il pourra être plus judicieux d'utiliser une base de fonctions non-périodiques si l'on sait que f est non-périodique, par exemple (ici $x \in [0,1]$) [7] :

$$u_k(x) = \cos \pi k x, \quad k = 0, 1, \ldots$$
$$v_k(x) = x^k, \quad k = 0, 1, \ldots$$
$$w_0(x) = 1,\ w_1(x) = x,\ w_k(x) = \sin \pi k x, \quad k = 2, 3, ..$$

Mentionnons également la possibilité d'utiliser les fonctions splines (polynômes par morceaux adéquatement raccordés). Nous renvoyons à [16].

Prenons un exemple : On s'intéresse à l'affluence dans des magasins (réponse y) en fonction du temps qu'il fait x (note combinant température et pluviosité). Le modèle le plus simple est

$$y_i = \beta_1 + \beta_2 x_i + e_i.$$

7. Comme pour les g_k, les combinaisons linéaires des f_k fonctions forment un ensemble dense dans l'espace des fonctions continues sur $[0,1]$ par application du théorème de Stone-Weierstrass ; les h_k étant essentiellement les primitives des f_k, on montre également la densité.
On peut préférer aux g_k des polynômes orthogonaux, ce qui théoriquement ne change rien mais pratiquement donne typiquement un meilleur conditionnement de $X^T X$.

Si les mesures sont prises à des heures différentes de la journée, il sera très important d'intégrer cela au modèle, par exemple par l'intermédiaire d'une variable t_i variant entre disons 9h et 19h :

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 u_0(\tilde{t}_i) + \cdots + \beta_{3+k} u_k(\tilde{t}_i) + e_i, \quad \tilde{t}_i = (t_i - 9)/10.$$

On a maintenant un modèle de régression habituel avec $k+3$ régresseurs (données longitudinales). On aurait pu également découper la journée en parties et introduire un régresseur catégorielle, ce qui revient au même que de prendre pour $u_0, ..u_k$ des fonctions indicatrices d'intervalle, et introduit des discontinuités assez peu naturelles.

Une autre façon de procéder pour fabriquer une base adéquate peut être de partir des données elles-mêmes (ou d'autres mesures), si elles s'y prêtent : si l'on dispose de mesures d'affluence prises dans différents magasins tout au long de la journée, $y_m(t_j)$, où $m = 1, \ldots M$ est l'indice de magasin et t_j est une suite d'instants donnés de l'intervalle $[9, 19]$, on peut faire une ACP de ces M vecteurs pour en extraire les composantes principales $u_i(t_j)$ dont les premières fourniront une base adaptée. Le fait d'utiliser les réponses pour fabriquer les variables explicatives va malheureusement complètement perturber les tests qui suivront.

La librairie GAM (Generalized Additive Models) de R identifie des modèles du type

$$g(E[y_i]) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

où les fonctions f_j sont estimées par des splines.

II.6 Régression sur des classes. Segmentation des données

Si l'hypothèse de linéarité n'est pas satisfaite, on a vu qu'une solution peut consister à ajouter de nouvelles variables explicatives basées sur les premières (logarithme, etc.)

L'option proposée ici est de faire une classification des données basée sur certaines variables explicatives puis faire une régression différente sur chaque classe. C'est une pratique assez courante sur les grands ensembles.

Ceci revient bien entendu à faire une régression globale avec de nouvelles variables explicatives tenant compte des classes, mais est en pratique plus simple à gérer.

II.7 Mélange de régressions

Ce modèle consiste à considérer que la loi de y est un mélange de gaussiennes gouverné par différents paramètres :

$$y \sim \sum_{r=1}^{R} p_r \mathcal{N}(x\beta_r, \sigma_r^2).$$

Ce qui peut également s'interpréter comme un mélange de plusieurs types de données, chacun suivant le modèle habituel ; chaque type r a la probabilité p_r d'être choisi. Dans le cas où le régresseur x ne contient que la constante, $x_i = 1$, on retrouve le mélange de gaussiennes. Le type n'est pas observé, c'est une variable latente.

Par exemple r peut représenter un certain type de consommateur, et y son opinion (note) sur un certain produit ; p_r est la proportion de consommateurs du type r. Ces types sont inconnus, et la régression permettra de les faire apparaître.

Ou encore X est la température de surface en mer sur une zone, y est le courant de surface (vecteur de $\mathbb{R}^3$) et r le type de mode dynamique [76].

Ce modèle avec $R = 2$, $\beta_1 = \beta_2$ et $\sigma_1 \neq \sigma_2$ a été utilisé pour modéliser des individus aberrants.

Mentionnons sans démonstration que l'estimation du modèle peut se faire itérativement par la méthode

EM, ce qui conduit aux équations de réestimation (p. ex. [55])

$$
\begin{aligned}
q_{ir} &\longleftarrow \frac{p_r G(y_i; x_i^T\beta_r, \sigma_r)}{\sum_s p_s G(y_i; x_i^T\beta_s, \sigma_s)} \\
p_r &\longleftarrow \frac{1}{n}\sum_{i=1}^{n} q_{ir} \\
\beta_r &\longleftarrow (X^T\Sigma_r X)^{-1} X\Sigma_r y, \quad \Sigma_r = Diag(q_{1r} \ldots q_{nr}) \\
\sigma_r^2 &\longleftarrow \frac{\sum_i q_{ir}(y_i - x_i\beta_r)^2}{\sum_i q_{ir}}
\end{aligned}
$$

où $G(y;\theta)$ désigne la densité gaussienne. Dans cet algorithme, q_{ir} représente la probabilité a posteriori (c-à-d après observation des réponses) que la donnée i soit du type r. Il faut bien entendu fournir des valeurs initiales.

II.8 Une remarque dans le cas de réponses vectorielles.

Si le modèle est

$$y_i = x_i\beta^* + e_i, \quad Var(e_i) = \Sigma_*^2$$

(y_i est un vecteur ligne) alors l'estimateur naturel de la matrice β^* sera

$$\arg\min_{\beta} \sum_i (y_i - x_i\beta)\Sigma_*^{-2}(y_i - x_i\beta)^T$$

qui correspond au maximum de vraisemblance gaussien lorsque Σ_* est connu. En pratique Σ_* est inconnu et l'on peut être tenté d'estimer simultanément β^* et Σ_* au maximum de vraisemblance, ce qui devient compliqué. C'est en fait inutile car la solution de l'équation ci-dessus est indépendante de Σ_* ! En effet, si $\widehat{\beta} = (X^TX)^{-1}X^TY$ est la solution obtenue en mettant la matrice identité et $\widehat{e}_i = y_i - x_i\widehat{\beta}$, on a pour tout autre β :

$$
\begin{aligned}
\sum_i (y_i - x_i\beta)\Sigma_*^{-2}(y_i - x_i\beta)^T &= \sum_i (\widehat{e}_i + x_i\widetilde{\beta})\Sigma_*^{-2}(\widehat{e}_i + x_i\widetilde{\beta})^T, \quad \widetilde{\beta} = \widehat{\beta} - \beta \\
&= \sum_i \widehat{e}_i\Sigma_*^{-2}\widehat{e}_i^T + 2\sum_i \widehat{e}_i\Sigma_*^{-2}\widetilde{\beta}^T x_i^T + \sum_i x_i\widetilde{\beta}\Sigma_*^{-2}\widetilde{\beta}^T x_i^T .
\end{aligned}
$$

Le terme médian est nul car il vaut $Tr(2\Sigma_*^{-2}\widetilde{\beta}^T\sum_i x_i^T\widehat{e}_i) = Tr(2\Sigma_*^{-2}\widetilde{\beta}^T X^T(Y - X\widehat{\beta}) = 0$, le premier est indépendant de β est le dernier est minimum pour $\beta = \widehat{\beta}$.

Tout se passe donc comme si l'on estimait chaque colonne de β_* indépendamment, chacune à partir des observations de la coordonnée de la réponse correspondante.

II.9 Surparamétrisation, réduction de modèle et sélection de variables

On est dans la situation où il y a un très (trop) grand nombre de régresseurs, voire même plus que d'individus : par exemple l'individu est un certain produit composé (un vin, etc.), le régresseur j est la mesure d'absorption de la lumière émise à une certaine fréquence ω_j et la réponse est la fraction d'un produit spécifique dans le composé [45]. Deux mesures correspondant à deux fréquences proches sont très corrélées. Si l'on s'intéresse à la présence de plusieurs produits il y a plusieurs réponses.

On considérera ici comme exemple un tableau X à 251 individus (des hydrocarbures) et 401 variables (un spectre) [8]. La figure II.4 montre les spectres pour 4 individus pris au hasard ; la réponse est la température

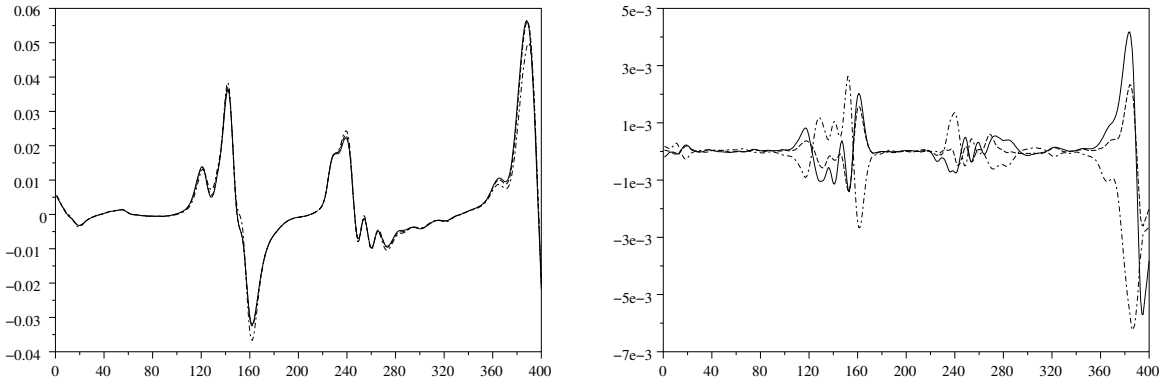


FIGURE II.4 – Spectres NIR de 4 échantillons d'hydrocarbure avant et après recentrage.

de gel. Dans les expériences qui vont suivre, on a recentré les variables et les réponses. Noter que l'on voit bien la corrélation des régresseurs en raison de la régularité en fréquence.

Aiji, Tavolaro, Lantz et Faraj [24] présentent un travail sur des données du même type avec 69 individus et 2232 variables explicatives (longueurs d'onde). P. Bastien [26] travaille sur 40 individus (sujets) et 1800 variables (taux d'expression de 1800 gènes).

Voici un exemple de situation à plusieurs réponses cité dans [3] : les réponses sont 6 caractéristiques d'un polymère en sortie d'un réacteur et les régresseurs sont 21 températures en divers endroits du réacteur et la vitesse de production ; il y a 26 individus. Autre exemple : les régresseurs sont des variables écologiques (activité humaine, caractérisations du milieu, etc.) et les réponses sont des mesures de densité d'espèces. On trouvera d'autres exemples dans [7], p peut aller à plusieurs dizaines de milliers.

Observations empiriques :

1. L'estimateur OLS a un faible pouvoir prédictif comparé à d'autres qui seront présentés plus bas.
2. Le surajustement est un problème récurrent (cf. § II.5).
3. Un petit sacrifice sur le RSS augmente considérablement le choix des β possibles : l'ensemble $B_\varepsilon = \{\beta : SS(\beta) \leqslant (1+\varepsilon)SS(\widehat{\beta})\}$ est volumineux.

Analyse : On est dans la même situation qu'au § II.5, la difficulté supplémentaire étant qu'il n'y a pas de relation d'ordre sur les coefficients, relation qui réduisait le problème à la comparaison d'une suite de modèles emboîtés. Le point 3 précédent encourage à aller dans la direction suivante : Introduire une hypothèse a priori sur β qui permettra essentiellement de choisir un point particulier dans B_ε, et proposer un estimateur dont les performances seront d'autant meilleures que cette hypothèse sera satisfaite. Le type d'hypothèses considéré sera « β de norme raisonnable » ou « β a peu de coefficients non nuls ».

Objectifs. On est conduit naturellement à des objectifs plus ou moins contradictoires dont l'importance dépendra de l'application considérée :

(i) *Pouvoir prédictif :* Trouver un estimateur (biaisé) de moindre MSE (et plus prédictif) sous une hypothèse a priori sur β.

(ii) *Parcimonie* (sparsity) : Obtenir beaucoup de β_j nuls sans affecter la prédiction.

(iii) *Sélection* : Estimer à 0 les β_j nuls (i.e. $\beta_j^* = 0$).

(iv) Traiter des situations pratiques intermédiaires où le nombre de variables explicatives est très grand et beaucoup de β_j sont petits sans être nuls.

Pour bien comprendre la littérature, il convient également de voir qu'il y a des situations où ce n'est pas la qualité d'une prédiction future qui importe (auquel cas c'est $\|X(\widehat{\beta} - \beta^*)\|$ qui est le critère) mais bien β lui-même (auquel cas c'est $\|\widehat{\beta} - \beta^*\|$ qui est le critère), comme en imagerie par scanner médical (tomodensitométrie), β est l'image et y une transformation linéaire de cette dernière (transformée de Radon : intégrales de β sur des droites). Dans le premier cas, l'inversibilité de X^TX n'est pas vraiment nécessaire, ce qui rend la situation différente.

8. Mis à librement disposition par Eigenvector Research, Inc., `www.eigenvector.com/data/index.htm` .

Les méthodes présentées dans la suite proposeront généralement une suite de modèles de complexité croissante. Le choix entre ces différents modèles reste délicat et se fait souvent avec la **validation croisée**.

Les estimateurs seront tous des fonctions de $\widehat{\beta}_{OLS}$ (et de X) ne faisant plus intervenir les réponses [9], ce qui se vérifie en notant qu'ils ne dépendent de y qu'au travers de sa projection sur $\mathcal{X}$.

On pourra consulter la référence [7] et le chapitre 3 de [11] pour des compléments à cette partie et examen des différents algorithmes récents.

II.9.1 Fabrication de nouveaux régresseurs par ACP ou PLS

Cette méthode est bonne si l'on ne cherche pas à se défaire du surajustement en sélectionnant des variables dont certaines seraient inutiles mais en éliminant la redondance qu'elles contiennent, comme dans l'exemple des hydrocarbures de la figure II.4, où beaucoup de variables se valent et où il ne s'agit probablement pas de choisir telle fréquence plutôt qu'une autre. Le modèle final utilisera toutes les variables.

Analyse en composantes principales. Une ACP de X transforme cette matrice en une matrice $X' = XW$ dont les colonnes sont orthonormées, les composantes principales, celles de plus grande inertie étant placées en premier.

Posons $X'_a = XW_a = X[w_1 \quad \dots \quad w_a]$. Les modèles proposés utilisent les a premières composantes :

$$y = X'_a\beta_a + u, \quad X'_a = (x'_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant a}.$$

On a $\widehat{\beta}_a = ({X'_a}^T X'_a)^{-1} {X'_a}^T y = ({X'_a}^T X'_a)^{-1} {X'_a}^T X\widehat{\beta}_{OLS}$, avec la prédiction sur un nouvel individu :

$$\widehat{y}_a = xW_a\widehat{\beta}_a.$$

On a donc un vecteur de régression $W_a\widehat{\beta}_a$, sur les variables originales.

Moindres carrés partiels (PLS). Vu l'objectif final, on peut trouver injuste que le calcul des composantes principales soit fait indépendamment des réponses ; d'où la méthode PLS, essentiellement utilisée dans le cas de plusieurs réponses, surtout à des fins de prédiction. L'idée est de choisir d'abord les combinaisons linéaires des régresseurs les plus corrélées aux réponses. Breiman et Friedman proposent dans [3] une discussion approfondie des différentes méthodes utilisées dans le cas de plusieurs réponses ; leurs conclusions sur le PLS sont plus que mitigées. L'appendice B présente l'algorithme et ses principes de base.

Attention, la validation croisée est, pour le PLS, lourde à mettre en œuvre du fait que les régresseurs sont calculés à partir des réponses. Il faut donc, pour éviter tout surajustement, retirer l'individu **avant** d'avoir commencé le PLS (ceci est moins critique pour l'ACP qui, elle, est faite sans utiliser les réponses), puis faire le PLS, la régression et calculer l'erreur de prédiction, ceci pour tous choix de nombre de composantes gardées et tous les individus. En sommant sur les individus on obtient un score de validation croisée pour chaque choix de nombre de composantes. C'est ce qui est fait dans la figure II.5 (figure de gauche) ; dans la figure de droite on a retiré non pas un individu mais 1/10 pris aléatoirement, ceci 60 fois, puis moyenné les erreurs obtenues (sur l'avantage de la V-fold CV, ici $V = 10$, voir l'appendice A). On compare avec l'ACP, l'ACP et la régression étant faites sur l'ensemble d'apprentissage.

Notons également que fait d'utiliser les réponses pour fabriquer les variables explicatives perturbe tout test de significativité que l'on pourrait faire ensuite, ce qui pousse à utiliser la validation croisée.

II.9.2 Ridge regression

Il ne s'agit plus ici de réduire le nombre de régresseurs mais de faire directement l'estimation en prenant en compte tous les régresseurs ; comme au § II.5, cette méthode a pour effet de réduire les coefficients de

9. D'un point de vue théorique, ceci une propriété que doit satisfaire tout estimateur digne de ce nom car $\widehat{\beta}_{OLS}$ est, *sous le modèle gaussien*, une statistique suffisante, et contient donc exactement toute l'information nécessaire.

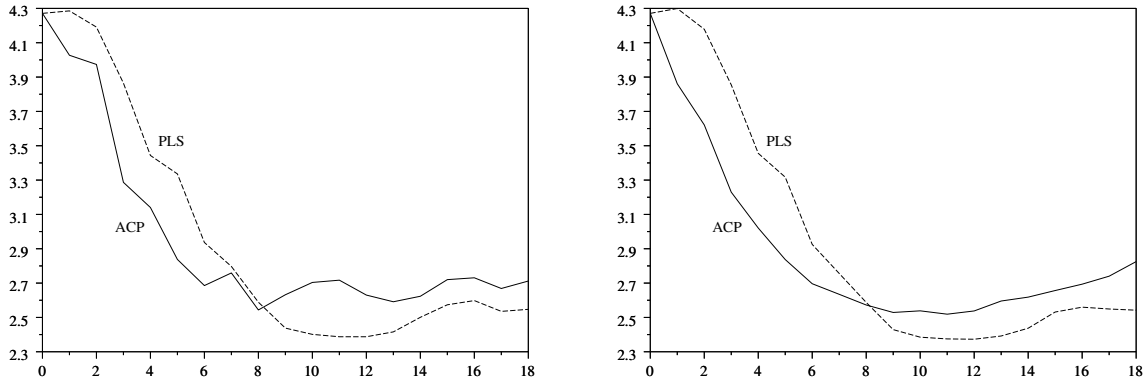


FIGURE II.5 – Critères de validation croisée sur les données d'hydrocarbure. Par leave-one-out (à gauche) et par extraction d'un paquet aléatoire de 10% d'individus test (à droite, V-Fold CV). En abscisse le nombre de régresseurs introduits, colonne de 1 non comprise ($\sqrt{TSS/n} = 4,27$).

$\widehat{\beta}_{OLS}$ («shrinkage»). La méthode de «ridge regression» propose l'estimateur suivant où le paramètre λ doit être estimé par validation croisée [44] :

$$\widehat{\beta}_R = (X^T X + \lambda Id_p)^{-1} X^T y. \tag{II.26}$$

C'est le β qui minimise le $SS(\beta)$ sous la contrainte que $\|\beta\| \leqslant \mu$ pour un certain μ (dépendant de λ). C'est également la solution de

$$\widehat{\beta}_R = \arg\min_{\beta} SS(\beta) + \lambda\|\beta\|^2. \tag{II.27}$$

Il est recommandé par défaut de normaliser les régresseurs. On peut vouloir préserver certaines colonnes de X (disons les premières) de l'effet d'atténuation, il suffit pour cela d'orthogonaliser les autres (aux premières) et de remplacer λId_p par une matrice diagonale ayant des 0 au début et des λ à la fin ; ainsi, lorsque λ tend vers l'infini on ne fait plus qu'une régression sur les premières colonnes. Sur les données d'hydrocarbure le meilleur λ choisi par validation croisée (leave one out) conduit à une valeur du critère de 2.39.

Arguments en faveur du ridge. Depuis l'estimateur de James-Stein, il est connu que la propriété BLUE (cf. p. 22) n'empêche pas qu'*il existe des estimateurs biaisés qui ont un meilleur MSE* et donnent de meilleurs prédicteurs. Ce phénomène s'accentue en grande dimension. Ici, un calcul simple montre que toujours, pour λ assez petit, l'estimateur ridge est meilleur que l'estimateur OLS au sens où $E[\|\widehat{\beta}_R - \beta^*\|^2] \leqslant E[\|\widehat{\beta}_{OLS} - \beta^*\|^2]$.

Si le plan est orthogonal, $X^T X = rId$, alors $\widehat{\beta}_R = \frac{r}{r+\lambda}\widehat{\beta}_{OLS}$ et le meilleur choix de λ est $\lambda_* = p\sigma_*^2/\|\beta^*\|^2$. Si $X = Id$ et si l'on estime $\|\beta^*\|^2$ dans cette dernière formule par $\|y\|^2 - n\sigma_*^2$, on retrouve l'estimateur de James-Stein $\widehat{\beta}_{JS} = (1 - p\sigma_*^2\|y\|^{-2})y$.

Le cas général peut être abordé en notant que la formule ridge correspond à $\widehat{\beta}_R = E[\beta^*|y]$ sous l'hypothèse bayésienne que $\beta^* \sim \mathcal{N}(0, \frac{\sigma_*^2}{\lambda} Id_p)$ (utiliser la formule classique pour les espérances conditionnelles de vecteurs gaussiens, note 5 p. 26). L'implication pratique de cela est que cet estimateur est *adapté au cas où les β_j^* (supposés i.i.d.) sont considérés comme étant a priori de même ordre de grandeur, et indépendants les uns des autres* ; une valeur raisonnable de λ est alors le rapport de σ_*^2 par la variance commune aux β_j^*. Il est donc important de tenter de *normaliser correctement les données* avant d'utiliser cet estimateur.

Exercice (Validation croisée) Vérifier que la formule (II.1) reste valide pour $\widehat{\beta}_R$ si l'on remplace $X^T X$ par $X^T X + \lambda Id$ (cf. l'exercice 13 p. 22) puis que l'erreur par validation croisée est donnée par (II.11) avec $h_i = x_i(X^T X + \lambda Id)^{-1} x_i^T$.

Le critère de validation croisée généralisée est $GCV(\lambda) = nRSS/(n - \mathrm{Tr}A_\lambda)^2$ où $A_\lambda = X(X^T X + \lambda Id)^{-1} X^T$ est la matrice telle que $A_\lambda y = \widehat{y}$ [44]. Vérifier que GCV coïncide avec CV si les h_i sont égaux, et qu'il généralise bien celui défini p. 33.

II.9.3 Méthodes récentes

On dispose à présent des méthodes suivantes :

1. *Forward selection* (méthode ascendante) : Elle consiste à créer une suite croissante de modèles en ajoutant à chaque étape la variable qui fait le plus diminuer le résidu $SS(\widehat{\beta})$ [10]. On élimine éventuellement à chaque étape des variables jugées désormais non-nécessaires (par un F-test, *stepwise selection*). Nous y reviendrons au § III.2.4.

   Elle est qualifiée de « quiet scandal in the statistical community » par Breiman [33]. Elle est peu stable au sens où le retrait de quelques échantillons peut modifier complètement le résultat [32]. Elle ne compare pas tous les modèles. C'est la fonction `stepAIC()` de R.
2. *Best subset regression* : Rechercher pour chaque k, le modèle à k régresseurs de moindre RSS. Sa mise en pratique est coûteuse [50]. Elle est peu stable, et donne des résultats moyens en prédiction lorsque l'on est dans la situation intermédiaire où beaucoup de β_j sont petits sans être nuls [32]. Fonctions `regsubsets()` ou `leaps()` de R, bibliothèque LEAPS. D'un point de vue théorique, elle est recommandée si le modèle est satisfait et si β^* a effectivement beaucoup de coefficients nuls.
3. *ACP et PLS.*
4. *Ridge* : Son pouvoir prédictif est meilleur que le best subset regression [32], mais il n'est pas invariant par changement d'échelle (sur les régresseurs), et ne conduit pas à des modèles plus simples (les $\widehat{\beta}_j$ sont $\neq 0$).

Notons que les méthodes ACP et ridge ont tendance à fusionner les régresseurs semblables ce qui est contraire à l'objectif de parcimonie présenté dans l'introduction, objectif qui pousse à choisir entre ces derniers. Les statisticiens se sont alors intéressé à proposer des méthodes intermédiaires entre best subset regression et ridge. L'idée est de considérer des généralisations de la méthode ridge de la forme suivante :

$$\widehat{\beta} = \arg\min_{\beta} SS(\beta) + \lambda \sum_j p(\beta_j), \quad SS(\beta) = \|y - X\beta\|^2 \tag{II.28}$$

où $p(.)$ est une certaine fonction de pénalisation et λ un paramètre. Noter que si l'on note $\varepsilon = SS(\widehat{\beta})$, $b = \sum_j p(\widehat{\beta}_j)$, on a clairement

$$\widehat{\beta} = \arg\min_{SS(\beta)=\varepsilon} \sum_j p(\beta_j) = \arg\min_{\sum_j p(\beta_j)=b} SS(\beta),$$

ce qui donne trois formulations équivalentes paramétrées par λ, ε, ou b. La seconde montre que $p(.)$ a pour but de choisir un point dans B_ε (cf. le point 3 p. 37) Noter aussi que les propriétés d'orthogonalité impliquent $SS(\beta) = \|X(\beta - \widehat{\beta}_{OLS})\|^2 + \|y - X\widehat{\beta}_{OLS}\|^2$ ce qui permet de voir que $\widehat{\beta}$ n'est fonction de y qu'au travers de $\widehat{\beta}_{OLS}$.

Méthode lasso. Il s'agit de l'estimateur correspondant à $p(x) = |x|$. Les colonnes de X sont généralement préalablement standardisées. Sur les données d'hydrocarbures, en utilisant le programme LARS disponible sous R et en choisissant λ par validation croisée, on trouve une valeur du critère de 2,5 avec 25 coefficients non nuls ; attention, ce chiffre est difficilement comparable aux 11 (en gros) variables choisies par le PLS ou l'ACP car il s'agit ici des variables originales. Le succès du lasso vient de ce que

1. Il est rapide à calculer, même en dimension élevée.
2. Il propose des solutions parcimonieuses (contrairement à l'estimée ridge).
3. Expérimentalement, ses performances sont comparables à celles de la sélection forward. L'écart est rarement flagrant, et dépend des situations.

Le chapitre 2 de [4], particulièrement les § 2.8 et § 2.9, présente clairement les variantes de mise en œuvre ; on voit qu'une deuxième étape est souvent introduite, p.ex. un OLS sur les variables conservées. L'« adaptive lasso » (qui est essentiellement le « garrote » de Breiman) semble en pratique souvent meilleur. Une mise en œuvre est faite pour les modèles linéaires généralisés par la bibliothèque GLMNET

10. Nous verrons au § III.2.4 qu'en fait, en présence de variables catégorielles, on minimise un critère qui tient également compte du nombre de paramètres ajouté.

sous R.

En comparaison avec la méthode stepwise, les simulations semblent montrer que lasso est plus stable (moins sensible aux fluctuations de l'échantillon) et plus adapté aux situations où il y a des petits coefficients ; la méthode best subset (en pratique une variante de stepwise) est plus adaptée aux situations où les coefficients de β_* sont effectivement bien concentrés sur un faible nombre de variables explicatives [81]. Il est sage en pratique de comparer les deux. La méthode forward reste tout-à-fait concurrentielle.

Le lasso est également en concurrence avec le «thresholded ridge» qui consiste à faire suivre le ridge d'une troncature à zéro des petits coefficients [74].

Méthode SCAD (Smoothly Clipped Absolute Deviation). Conçue pour pallier un biais du lasso (le lasso a tendance à raboter tous les coefficients, ce qui s'explique simplement à partir de la formule (II.28) avec $p(x) = |x|$), elle correspond à un choix de $p(.)$ plus compliqué que lasso, ressemblant qualitativement à $x \mapsto \min(|x|, 2\lambda)$, et son caractère non-convexe rend le calcul de $\widehat{\beta}$ moins évident. Elle semblerait jouir de bonnes propriétés lorsque $p \gg n$ [56] mais ceci est très discuté. La méthode $MC+$ (cf. p.ex. [62]), analogue, lui semble préférée.

Méthode Elastic net. Cette méthode réalise un continuum entre ridge et lasso en combinant les pénalisations. Le meilleur point du continuum étant choisi par validation croisée. En pratique elle fera donc généralement, en prédiction, mieux que les deux.

Screening. Il s'agit d'opérer une première passe assez simple pour éliminer des variables, particulièrement dans le cas $p \gg n$, avant d'utiliser une des méthodes précédentes, qui ne sont pas très bonnes pour p très grand [41]. L'idée la plus répandue est simplement d'ordonner les variables en fonction de leur corrélation avec la réponse ; elle est naïve mais concurrentielle si p est réellement très grand car peu gourmande en calcul ; nous renvoyons au § 4 de [7] et à [63]. Son avantage est, intuitivement, que l'évaluation de la contribution d'un régresseur ne sera ensuite pas perturbée par l'introduction hâtive d'autres régresseurs douteux dans le modèle, comme cela peut arriver avec la méthode forward. Le screening peut également se faire sur les composantes principales (§ 16.6 de [11]).

Surajustement. On a vu au § II.5 que si p est proche de n, le surajustement peut avoir des conséquences dramatiques, décelés par la validation croisée. Si maintenant $p \gg n$, il y a une grande quantité de modèles qui interpolent les données ($y = X\widehat{\beta}$) et il se peut que certains d'entre eux soient raisonnables ; on observe que la régression ridge avec $\lambda = 0$, qui choisit parmi tous ces modèle celui de $\widehat{\beta}$ de norme minimale ($\widehat{\beta} = (X^T X)^+ X^T y$, où M^+ est la pseudo-inverse de M) peut donner de bon résultats (cet estimateur est parfois l'estimateur OLS fourni par le logiciel dans le cas $p > n$). Il ne faut donc pas être surpris si la validation croisée en estimation ridge conduit à un pénalisation nulle.

II.9.4 Régression à rang réduit. Curds and whey

Ces deux méthodes concernent la situation où il y a de nombreuses réponses. La régression à rang réduit se propose de minimiser $\|Y - X\beta\|$ (norme de Frobenius) sous la contrainte que β a son rang inférieur à r donné. On obtient les étapes de calcul suivantes

$$\begin{aligned}
&\widehat{Y} = X\widehat{\beta}_{OLS} \\
&Q = \widehat{Y}^T\widehat{Y}(Y^TY)^{-1} = Y^TX(X^TX)^{-1}X^TY(Y^TY)^{-1} \\
&Q = T^{-1}DT \quad \text{(diagonalisation)} \\
&\widehat{\beta} = \widehat{\beta}_{OLS}T^{-1}I_rT
\end{aligned} \tag{II.29}$$

(c-à-d T est la matrice de passage de la diagonalisation de Q), et I_r est la matrice où les r plus grands coefficients de D sont mis à 1 et les autres à 0. r s'estime par validation croisée.

Breiman et Friedman [3] proposent une méthode («curds and whey») basée sur un modèle où X a une structure aléatoire de vecteurs indépendants. Sans entrer dans les détails, mentionnons qu'il obtient les nouvelles estimées comme combinaisons linéaires des estimées OLS pour chaque composante avec la formule

$$\widehat{\beta} = \widehat{\beta}_{OLS}((1-\lambda)I + \lambda Q^{-1})^{-1} = \widehat{\beta}_{OLS}T^{-1}((1-\lambda)I + \lambda D^{-1})^{-1}T$$

(La dernière formule permet de faire apparaître l'analogie avec (II.29) par comparaison des fonctions $1_{x>\rho}$ et $(1-\lambda+\lambda x^{-1})^{-1}$, $0 \leqslant x \leqslant 1$, $\lambda \sim \rho$). Le paramètre λ doit être estimé par validation croisée.

II.10 Régression robuste

Il s'agit de résister aux individus aberrants. On estime β^* par

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \sum \rho(y_i - x_i\beta)$$

où ρ est maintenant une certaine fonction différente du carré. Pour donner moins de poids aux individus aberrants, on choisit une fonction à croissance moins rapide que x^2. Typiquement

$$\rho(u) = u^2 1_{|u|\leqslant\alpha} + (2\alpha|u| - \alpha^2)1_{|u|>\alpha}.$$

Cette fonction vaut u^2 pour $|u|$ petit, et est d'ordre $|u|$ ensuite. Ce choix est justifié par des arguments théoriques précis dûs à Huber [52]. Voir [23] pour les détails pratiques. Quand $\alpha = +\infty$ on retrouve la méthode habituelle, et quand α tend vers 0, $\rho(u)/\alpha$ tend vers $2|u|$, et l'on retrouve la « régression l_1 » :

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \sum |y_i - x_i\beta|.$$

Exercice. Montrer que l'estimation de β^* en régression robuste peut s'interpréter comme le maximum de vraisemblance sous le modèle habituel, mais en modifiant la distribution de u.

III

Régression linéaire gaussienne, diagnostic et tests

III.1 Propriétés statistiques fondamentales des estimateurs

III.1.1 Modèle statistique et estimateurs

C'est le même que celui de la section II.2 sauf que les u_i sont supposés gaussiens :

Modèle. *On suppose l'existence d'un vecteur β^*, de $\sigma_* > 0$ tels que*

$$y \sim \mathcal{N}(X\beta^*, \sigma_*^2 I).$$

En d'autres termes, $y_i = x_i\beta^* + u_i$ et les u_i sont indépendants gaussiens centrés de variance σ_*^2. En pratique on pourra conforter cette hypothèse par un test de normalité sur les $\widehat{u}_i$ (p. ex. un test de Shapiro).

$\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}^2)$ est un estimateur sans biais de (β^*, σ_*^2), cf. § II.2.2. On va voir que la variance de $\widehat{\sigma}^2$ est $2\sigma_*^4/(n-p)$. L'estimateur au maximum de vraisemblance de θ^* est $\widehat{\theta}_{MV} = (\widehat{\beta}, \frac{n-p}{n}\,\widehat{\sigma}^2)$.

III.1.2 Propriétés de base des variables gaussiennes

Ce paragraphe a pour but de rappeler des propriétés élémentaires des vecteurs gaussiens. On ne détaillera pas les démonstrations.

13 - Définition

U est un vecteur aléatoire gaussien sur $\mathbb{R}^n$ ssi pour tout vecteur $a \in \mathbb{R}^n$, la variable aléatoire $\langle a, U\rangle$ est gaussienne.

On peut montrer que si la matrice de covariance R de U est définie positive, U a une densité (par rapport à la mesure de Lebesgue) qui s'exprime en fonction de la moyenne μ et de R comme suit :

$$p(u) = \frac{1}{\sqrt{(2\pi)^n \det(R)}} \exp\left\{-\frac{1}{2}(u-\mu)^T R^{-1}(u-\mu)\right\}.$$

Une des propriétés fondamentales des vecteurs gaussiens est l'équivalence entre indépendance et non-corrélation (elle est fausse pour les vecteurs de variables gaussiennes, cf. l'exercice 2 p. 45) :

14 - Théorème

Soient V et W deux vecteurs aléatoires tels que $U = \begin{pmatrix} V \\ W \end{pmatrix}$ forme un vecteur gaussien. Si V et W sont décorrélés, alors ils sont indépendants.

Ce résultat est simple à vérifier si $R > 0$ car l'hypothèse sur U implique que R est bloc-diagonale avec un bloc correspondant à V et un bloc correspondant à W ; il s'ensuit que la densité de U se factorise en $p(u) = p_1(v)p_2(w)$, ce qui implique l'indépendance.

Il est bon de voir que les vecteurs gaussiens s'expriment toujours comme combinaisons de v.a. gaussiennes indépendantes :

15 - Proposition

Soit $U \sim \mathcal{N}(\mu, R)$ de dimension n, alors il existe un vecteur gaussien centré réduit $V \sim \mathcal{N}(0, I)$ de dimension $m = \text{rang}(R)$ et une matrice Σ de dimension $n \times m$, tels que

$$U = \Sigma V + \mu, \qquad \Sigma\Sigma^T = R.$$

Si R est inversible prendre par exemple $\Sigma = R^{-1/2}$ (racine carrée symétrique) et $V = \Sigma^{-1}(U - \mu)$; sinon écrire $R = PDP^T$ où P est $n \times m$ avec $P^TP = I$ et D diagonale positive, puis poser $V = D^{-1/2}P^T(U - \mu)$, $\Sigma = PD^{1/2}$ (remarquer que $Var((I - PP^T)U) = 0$).

On utilisera fortement dans la suite la proposition suivante dont la démonstration est l'application des résultats précédents (cf. exercice 3 p. 45) :

16 - Proposition

Soit $U \sim \mathcal{N}(\mu, Id)$. Soient A_i, $i = 1, \dots q$, des matrices de projection orthogonale de dimension n. Alors

- Si $A_iA_j = 0$ pour tous $i \neq j$, alors les variables A_iU sont indépendantes et donc également les U^TA_iU.
- Si $\mu = 0$, alors U^TA_iU suit une loi de χ^2 à $r = \text{rang}(A_i) = \text{trace}(A_i)$ degrés de liberté.

En particulier si $U \sim \mathcal{N}(\mu, \sigma^2 Id)$, alors deux projections de U sur deux espaces orthogonaux sont indépendantes.

III.1.3 Loi de probabilité des estimateurs

On peut passer maintenant aux conséquences pour les estimateurs :

17 - Théorème

Sous l'hypothèse $y \sim \mathcal{N}(X\beta^*, \sigma_*^2 I)$:

- $\widehat{\beta} \sim \mathcal{N}(\beta^*, \sigma_*^2(X^TX)^{-1})$
- $(n - p)\ \frac{\widehat{\sigma}^2}{\sigma_*^2} \sim \chi^2_{n-p}$.
- $\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendants.

Démonstration. En effet $\widehat{\beta} = \beta^* + (X^TX)^{-1}X^Tu$ et $\widehat{\sigma}^2 = (n-p)^{-1}\|Ku\|^2$ (notations de la proposition 5). L'indépendance vient de la décorrélation de X^Tu et Ku. ∎

On montre également que $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}^2)$ est un estimateur de variance minimale dans la classe des estima-

teurs sans biais.

III.1.4 Exercices

Exercice 1. Vérifier que la vraisemblance de l'échantillon après estimation au maximum de vraisemblance, i.e. la densité de la gaussienne de moyenne $X\widehat{\beta}$ et de variance $\widehat{\sigma}^2_{MV} I$ appliquée à $(y_1, \dots y_n)$, est

$$p(y_1, \dots y_n) = (2\pi e \widehat{\sigma}^2_{MV})^{-n/2}.$$

Exercice 2. Soit U une variable $\mathcal{N}(0,1)$ et X un jeu de pile ou face équiprobable ($P(X=1) = P(X=-1) = 1/2$) indépendant de u. Montrer que $V = XU$ est $\mathcal{N}(0,1)$ et que U et V sont décorrélés mais pas indépendants (pour ce dernier point on pourra calculer $E[U^2V^2]$).

Exercice 3. Le but de cet exercice est la démonstration de la proposition 16. On rappelle que les matrices de projection orthogonale sont exactement les matrices symétriques P telles que $P^2 = P$.

1. Démontrer le premier point.
2. (a) Montrer que si $U \sim \mathcal{N}(0, Id)$ est Q est une matrice orthogonale, alors $QU \sim \mathcal{N}(0, Id)$.
   (b) En déduire le dernier point en diagonalisant A_i (ses valeurs propres valent toutes 0 ou 1).

Exercice 4. On se donne le modèle $y = X\beta^* + u$ où les u_i sont i.i.d de loi de densité $e^{-|u|/\lambda^*} du/(2\lambda^*)$. Exprimer la vraisemblance des observations pour une paire donnée (β, λ), et donner l'expression de l'estimateur au maximum de vraisemblance de λ quand β^* est connu.

III.2 Analyse de l'estimateur

III.2.1 Détermination d'intervalles de confiance

Rappelons que la loi de Student de paramètre k est celle de $X/\sqrt{Y/k}$ où X est une gaussienne centrée réduite et Y un χ^2_k indépendant. La loi de Fisher-Snedecor (k,l) est celle de $(X/k)/(Y/l)$ où $X \sim \chi^2_k$ et $Y \sim \chi^2_l$ sont indépendants. On désignera par $t_k(.)$ et $f_{kl}(.)$ les fonctions quantile de ces distributions.

Comme conséquence immédiate du théorème 17 et de la proposition 16, on a les propriétés suivantes

18 - Proposition

Sous l'hypothèse $y \sim \mathcal{N}(X\beta^*, \sigma^2_* I)$:

- Pour tout $j = 1, \dots p$, la variable aléatoire

$$T_j = \frac{\widehat{\beta}_j - \beta^*_j}{\widehat{\sigma}(\widehat{\beta}_j)}$$

  suit une loi de Student de paramètre $n-p$ ($\widehat{\sigma}(\widehat{\beta}_j)$ est l'erreur standard de $\widehat{\beta}_j$, cf. § II.2.2).
- Pour tout vecteur u, la variable aléatoire

$$T_u = \frac{u^T\widehat{\beta} - u^T\beta^*}{\widehat{\sigma}(u^T\widehat{\beta})}, \quad \widehat{\sigma}(u^T\widehat{\beta})^2 = \widehat{\sigma}^2 u^T (X^T X)^{-1} u$$

  suit une loi de Student de paramètre $n-p$.
- Soit $q < p$ et L une matrice $q \times p$ de rang q, la v.a

$$F = \frac{1}{q\widehat{\sigma}^2} (\widehat{\beta} - \beta^*)^T L^T \left(L(X^TX)^{-1}L^T\right)^{-1} L(\widehat{\beta} - \beta^*)$$

  suit une loi de Fisher-Snedecor de paramètres $(q, n-p)$.

Les deux premiers points sont une conséquence du théorème 17 ; pour le troisième, noter que la variable $\left(L(X^TX)^{-1}L^T\right)^{-1/2} L(\widehat{\beta}-\beta^*)$ suit la loi $\mathcal{N}(0,\sigma_*^2 Id_q)$, ce qui fait que le numérateur est un σ_*^2 fois un χ_q^2.

Un intervalle de confiance. En raison de la symétrie de la loi de Student on a $P(|T_j| < t_{n-p}(1-\alpha/2)) = 1-\alpha$. On obtient donc un intervalle de confiance *de probabilité de confiance* $1-\alpha$ *pour le coefficient* β_j^*

$$\left[\widehat{\beta}_j - \delta\ ,\ \widehat{\beta}_j + \delta\right], \quad \delta = \widehat{\sigma}(\widehat{\beta}_j)\ t_{n-p}(1-\alpha/2) \tag{III.1}$$

Une région de confiance. De la même façon la relation $P(F < f_{q,n-p}(1-\alpha)) = 1-\alpha$ se réécrit $P(L\beta^* \in \mathcal{R}_\alpha) = 1-\alpha$ où

$$\mathcal{R}_\alpha = \left\{\zeta \in \mathbb{R}^q : \|L\widehat{\beta} - \zeta\|^2_{[L(X^TX)^{-1}L^T]^{-1}} \leqslant q\widehat{\sigma}^2 f_{q,n-p}(1-\alpha)\right\}$$

(on note $\|x\|_S = x^TSx$) qui est donc une région de confiance de probabilité de confiance $1-\alpha$ pour le vecteur $L\beta^*$.

On obtient une région de confiance pour $(\beta^*_{j_1}, \dots \beta^*_{j_q})$ si L est la matrice de sélection $q \times p$ telle que $L\beta = (\beta_{j_1}, \dots \beta_{j_q})$.

III.2.2 Rappels sur les tests dans le cadre paramétrique général

On supposera que l'on a un modèle paramétrique P_θ, $\theta \in \Theta$, pour un ensemble de données $Y = (y_1, \dots y_n)$, et que l'on cherche à décider entre H_0 : « $\theta^* \in \Theta_0$ » et H_1 : « $\theta^* \in \Theta_1$ », avec $\Theta_0 \cap \Theta_1 = \varnothing$ (dans le cas général non paramétrique H_0 et H_1 sont deux ensembles de lois de probabilité candidates pour Y).

On s'intéressera en particulier au cas où H_1 =non-H_0, i.e. $\Theta_0 \cup \Theta_1 = \Theta$; si H_0=« $\theta^* = \theta_0$ » (Θ_0 réduit à un singleton) on dit que H_0 est simple. L'idée est que celui qui met le test en œuvre cherche à convaincre de la véracité de H_1.

Un test $\varphi = \varphi(Y) \in \{0,1\}$ décidant entre les hypothèses H_0 et H_1 est de **niveau** α (petit) ssi :

toujours sous H_0, $P(\varphi = 1) \leqslant \alpha$.

La probabilité d'erreur de première espèce (choisir H_1 à tort) est au plus égale à α. Un faible niveau est donc seulement une garantie que H_1 **sera acceptée à bon escient**. Par exemple le test qui choisit systématiquement H_0 a un niveau égal à zéro (mais aucun intérêt). L'importance du niveau s'illustre par l'exemple type où H_0 est « Ce médicament est sans effet » et H_1 : « Ce médicament a un effet positif » ; il est clairement important de ne pas décider H_1 si H_0 est vraie (mise sur le marché d'un médicament sans effet) ; d'où le terme de **test de significativité**. Même remarque pour H_0 : « Le diesel et le sans plomb sont aussi polluants » et son contraire.

> Si en revanche le test décide H_0, c'est sa puissance qui permet de conclure : On dit que le test est de **puissance** $1-\beta$ (proche de 1) ssi la probabilité d'erreur de deuxième espèce est inférieure à β :
>
> toujours sous H_1, $P(\varphi = 0) \leqslant \beta$.
>
> Ce concept n'est pas d'une grande aide pour les tests d'une hypothèse contre son contraire, car la puissance vaut alors typiquement α : l'ensemble H_1 contient des distributions arbitrairement proches de H_0 ; sous ces distributions, le test décidera H_0 avec probabilité au moins $1-\alpha$, ce qui implique puissance inférieure à α. On dit que le test φ est **plus puissant** que φ' si :
>
> toujours sous H_1, $P(\varphi = 1) \geqslant P(\varphi' = 1)$.
>
> On dit que le test φ est UPP (universellement plus puissant) s'il est plus puissant que tout autre test de même niveau. C'est ce type de test qui est recherché quand H_0 et H_1 contiennent des hypothèses arbitrairement proches. Pour revenir à l'exemple du médicament, l'organisme payeur (Sécurité Sociale) veut un niveau faible garanti tandis que le laboratoire veut un test puissant. Ces deux exigences ne peuvent être conciliées qu'avec un minimum d'échantillons.

Mise au point d'un test d'hypothèses. La méthode usuelle consiste à utiliser une statistique $S(Y)$ dont la valeur est plutôt faible sous H_0 et grande sous H_1 (p.ex. $S(Y) = \|\widehat{\theta}\|$ si H_1 : « $\theta^* \neq 0$ », $S(Y) = \widehat{\theta}$ si H_1 : « $\theta^* \geqslant 0$ ») et à rejeter H_0 si $S(Y)$ est trop grand :

1. Choisir une statistique $S(Y)$ dont la loi est toujours la même sous H_0 (statistique « pivotale ») ; la puissance sera d'autant meilleure que $S(Y)$ sera grande sous H_1.
2. Se donner un niveau α.
3. Rejeter l'hypothèse si la valeur de S est déraisonnablement grande :

Rejeter H_0 si $\quad S(Y) > Q_S(1-\alpha)$

   où Q_S est la fonction quantile de S sous H_0.

   - **Rejet** : « la valeur de la statistique écarte H_0 (au niveau α) ; choisir H_1 ». La p-value, valeur de α telle que $S(Y) = Q_S(1-\alpha)$ est une mesure de la confiance de l'on doit porter au résultat.
   - **Acceptation** : « rien ne permet d'invalider H_0 ». Ex. : on n'a pas pu prouver l'efficacité du médicament. Si la la puissance est faible (ce qui est souvent le cas), H_0 *peut très bien être fausse* ; ceci se produit si l'on manque de données, ou si le test est peu performant, ou si la vraie loi est proche de H_0.

Si H_0 est simple, toute statistique est pivotale. Souvent cependant H_0 est multiple et l'on ne dispose pas de statistique pivotale ; pour garantir un niveau α, on sera contraint de faire un test de la forme : « Rejeter H_0 si $\quad S(Y) > \max Q_S(1-\alpha)$ » où le maximum est pris sur toutes les distributions de H_0.

Dans de nombreuses situations pratiques, la loi de S sous H_0 n'est connue qu'asymptotiquement (infinité d'observations). Dans ce cas on pourra avoir avantage à estimer $Q_S(1-\alpha)$ sous H_0 par des simulations ; si le modèle est paramétrique, et H_0 est simple H_0 : « $\theta^* = \theta_0$ » :

1. Simuler N jeux de données $Y^1, \dots Y^N$ sous θ_0 (choisir $N \gg 1/\alpha$)
2. On estime alors $Q_S(1-\alpha)$ par un réel séparant les αN plus grandes valeurs de $S(Y^i)$ des $(1-\alpha)N$ plus petites.

Si maintenant H_0 n'est pas simple, on peut reprendre cette méthode en utilisant cette fois-ci $\widehat{\theta}_0$, l'estimée de θ^* sous H_0 pour faire les simulations (...au lieu de faire le max sur tous les θ de H_0 comme il faudrait en toute rigueur). Cette méthode peut se justifier mathématiquement si la statistique S est asymptotiquement pivotale.

P-value. On ne donne en général pas le résultat ddu test mais la p-value, solution de $S(Y) = Q_S(1-\alpha)$, c'est-à-dire le seuil critique. C'est la probabilité sous H_0 d'observer une valeur de S au moins aussi grande.

Le seuil de 5%. La valeur critique de 5% et souvent utilisée. Elle est extrêmement élevée car elle autorise à mentir une fois sur 20, c'est-à-dire beaucoup plus souvent s'il l'on oublie de faire état d'expériences conduisant à une acceptation de H_0 (p.ex. en réitérant une expérience jusqu'à l'obtention de H_1, ce qui n'exige en gros qu'une vingtaine de répétitions, mais est affreusement non-déontologique). Cette valeur a été introduite par Fisher et présentée en ces termes ([42] p.82) :

> *"If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05 ..."*

Il ne s'agit donc à l'origine que d'un valeur indicative raisonnable pour lever une suspicion [64].

Dualité tests/intervalles de confiance. On estime un paramètre vectoriel θ caractérisant la loi des données. Si l'on dispose d'un domaine aléatoire I (typiquement de la forme $I = \{\theta : |\theta - \widehat{\theta}| \leqslant \delta\}$) tel que pour tout θ, $P_\theta(\theta \in I) \geqslant 1-\alpha$, alors le test φ_I qui accepte l'hypothèse « $\theta^* = \theta_0$ » ssi $\theta_0 \in I$ a un niveau inférieur à α (vérification immédiate). Notons que son niveau et sa puissance sont fonction croissante de I (car $\varphi_I \leqslant \varphi_{I'}$ si $I \subset I'$).

Réciproquement si l'on dispose pour tout θ_0 d'un test $\varphi(\theta_0)$ de niveau au plus α entre H_0 : « $\theta^* = \theta_0$ » et H_1 : « $\theta^* \neq \theta_0$ », alors l'ensemble aléatoire

$$I = \{\theta_0 : H_0 \text{ est acceptée}\}$$

est dit région de confiance de probabilité de confiance $1-\alpha$ et

$$P_{\theta_0}(\theta_0 \in I) = P_{H_0}(\varphi = 0) \geqslant 1-\alpha.$$

Exemple : Test de nullité d'un coefficient β_j^*. En vertu de la proposition 18, la statistique $T = |\widehat{\beta}_j|/\widehat{\sigma}(\widehat{\beta}_j)$ est pivotale pour H_0 : « $\beta_j^* = 0$ » (la loi de T sous H_0 est indépendante de σ_* et des autres β_k^*), et l'on a le test de niveau α pour décider que β_j^* est significativement différent de zéro (cf. § III.1) :

$$\frac{|\widehat{\beta}_j|}{\widehat{\sigma}(\widehat{\beta}_j)} \gtrless t_{n-p}(1-\alpha/2).$$

On présente souvent les résultats d'une régression avec un tableau contenant les niveaux de signification de ces hypothèses (test de type III, procédure SUMMARY de R ou S+). Donnons par exemple en table III.1 le cas de la prédiction de la consommation des voitures [49] en fonction des variables « Volume », « Puissance » et « Poids » (un tracé réponses/régresseurs fait préférer la variable P2 qui est le carré du poids) :

	Estimée $\widehat{\beta}_j$	Écart-type $\widehat{\sigma}(\widehat{\beta}_j)$	t-stat	p-value
Volume	-7,2 e-6	1,7 e-5	-0,42	0,67
P2	1,5 e-5	1,26 e-6	12	$<$ 2 e-16
Puissance	4,6 e-5	1,13 e-5	4,1	0,0001

TABLE III.1 – Table d'analyse des coefficients (82 individus). La colonne t-stat contient la statistique de student, rapport des deux premières colonnes.

INTERPRÉTATION : On va voir au paragraphe suivant que α_j est directement lié à la différence de RSS entre le modèle original et le modèle sans la j-ième variable explicative. α_j s'interprète donc également comme une mesure de l'amélioration de la prédiction due à l'introduction du j-ième régresseur après tous les autres. Un grand α_j ne signifie cependant pas que les réponses sont (presque) indépendantes du j-ième régresseur, car ce dernier peut être fortement corrélé aux autres, c'est le problème des *facteurs proches* (ou encore de la *colinéarité*) : Si dans l'exemple on rajoute comme régresseur la vitesse maximale de la voiture, la puissance semble alors non significative :

	Pr
Volume	0,63
P2	$<$ 0,0001
Puissance	0,83
Vitesse	0,7

Même si sur certaines données on doute de la validité du modèle (résidus gaussiens...) et que l'on ne désire pas interpréter les p-values α_j au pied de la lettre, ces dernières peuvent toujours être considérées comme des *instruments de mesure* qui résument au mieux l'information d'intérêt, au sens où leur calcul respecte toutes les règles de normalisation fondamentales déduites du modèle gaussien.

Au vu de la formule (II.3), on voit que *chaque test de student est un test de corrélation partielle entre la réponse et la variable explicative sachant les autres.*

III.2.3 Test de Fisher

On vient de voir comment le test de nullité d'un coefficient permet d'étudier la significativité d'une variable explicative. Malheureusement, si cette variable est catégorielle elle interviendra dans plusieurs coefficients ; il faut donc être capable de tester la nullité simultanée de plusieurs coefficients. On est donc conduit à tester H_0 : « $L\beta^* = l$ » où $l = 0$ et L est ici une matrice dont chaque ligne contient exactement un 1 et $p-1$ zéros. C'est ce que font les tests de type 1 et 3 des logiciels par opposition aux tables d'analyse des coefficients qui considèrent chaque modalité séparément (nous y reviendrons au § III.4.3).

Test de Fisher. Soit $L \in \mathbb{R}^{q\times p}, l \in \mathbb{R}^q$, le test de Fisher de niveau α pour H_0 : « $L\beta^* = l$ » est $l \in \mathcal{R}_\alpha$, soit

$$\|L\widehat{\beta} - l\|^2_{[L(X^TX)^{-1}L^T]^{-1}} \gtrless q\widehat{\sigma}^2 f_{q,n-p}(1-\alpha)$$

Il existe une réécriture du membre de gauche qui s'avère très utile (encadré ci-dessous) ; elle est basée sur le lemme suivant démontré à l'exercice 11 p. 22 :

19 - Lemme

Soit $\widehat{\beta}_0$ l'estimateur de β^* aux moindres carrés sous la contrainte $L\beta = l$, et $\widehat{y}_0 = X\widehat{\beta}_0$. On a

$$(L\widehat{\beta} - l)^T \left(L(X^TX)^{-1}L^T\right)^{-1} (L\widehat{\beta} - l) = \|\widehat{y} - \widehat{y}_0\|^2 = \|\widehat{y}_0 - y\|^2 - \|\widehat{y} - y\|^2 = RSS_0 - RSS.$$

Le test de Fisher présenté plus haut équivaut donc à

$$\text{Rejeter } H_0 \text{ si } \quad \frac{(RSS_0 - RSS)/(p - p_0)}{RSS/(n-p)} > f_{q,n-p}(1-\alpha)$$

où RSS_0 est le résidu calculé sous H_0, et $p_0 = p - q$ le nombre de paramètres du modèle sous H_0. Ce test généralise les tests de Student du paragraphe précédent (et les écrits d'une façon différente). Il est simple de vérifier que cette statistique est fonction de la *statistique du rapport de vraisemblance généralisé* (i.e. $(RSS_0/RSS)^n$, cf. exercice 8 p. 53), ce qui lui donne une interprétation naturelle. Il peut se justifier directement par la remarque suivante :

$$\text{Sous } H_0\text{, les statistiques } \frac{RSS_0 - RSS}{\sigma_*^2} \text{ et } \frac{RSS}{\sigma_*^2} \text{ sont des } \chi^2_q \text{ et } \chi^2_{n-p} \text{ indépendants}$$

en vertu du théorème de Cochran (proposition 16). Le numérateur de la statistique de Fisher vaut également $ESS - ESS_0$ et peut s'interpréter comme un terme dû à l'écart entre H_0 et H_1 dans la décomposition de la variance

$$TSS = (ESS - ESS_0) + ESS_0 + RSS$$

la statistique ne faisant que mesurer l'importance relative de cette partie de variance expliquée. Sous H_0 ces trois termes sont indépendants, nous en laissons la vérification en exercice.

Ce test possède des propriétés d'optimalité et d'invariance que nous ne détaillerons pas ici (cf. p. ex. [71] p. 46).

Le test de Fisher avec une matrice L générale peut être réalisé sous R avec la commande `lht` de la bibliothèque `car`.

Interprétation : Soit $F = \frac{(RSS_0 - RSS)/(p-p_0)}{RSS/(n-p)}$ la statistique de Fisher. On présente le résultat du test en donnant la valeur critique du seuil $\alpha = 1 - F_{p-p_0,n-p}(F)$, où $F_{p-p_0,n-p}$ est la fonction de répartion de la loi de Fisher-Snedecor de paramètres $p-p_0$ et $n-p$ (comme à la table III.1). L'hypothèse H_0 : « $L\beta^* = l$ » est refusée si α est inférieur au niveau α_0 (par exemple $\alpha_0 = 5\%$), c-à-d si F appartient à un intervalle de $[f_\alpha, +\infty[$ de probabilité 5% sous H_0, c-à-d, est anormalement grande sous H_0 (RSS petit).

Test de nullité de β^*. On teste la nullité des coefficients d'indice supérieur ou égal à 2 :

$$\frac{(n-p)ESS}{(p-1)RSS} \geqslant f_{p-1,n-p}(1-\alpha).$$

La **table d'analyse de variance** (ANOVA) résume la situation sous un format traditionnel où est donnée la valeur critique de α :

	SS	d.l.	F-stat	Pr
Modèle	ESS	$p-1$	F	$1 - f_{p-1,n-p}(F)$
Résidu	RSS	$n-p$		
Total	TSS	$n-1$		

$$F = \frac{(n-p)ESS}{(p-1)RSS}$$

L'hypothèse H_0 :« $\beta_i = 0, i \geqslant 2$ » est refusée si Pr est inférieur au niveau α (par exemple $\alpha = 5\%$). La colonne d.l. contient les degrés de liberté des statistiques (SS) qui sont des χ^2 sous H_0.

Test de nullité partielle de niveau α pour décider si $(\beta^*_{j_1}, \dots \beta^*_{j_q})$ est significativement non-nul :

$$\frac{(RSS_q - RSS)/q}{RSS/(n-p)} \geqslant f_{q,n-p}(1-\alpha). \tag{III.2}$$

où RSS_q est l'erreur résiduelle du modèle estimé sous la contrainte que les $\beta^*_{j_i}$ sont nuls. L est ici la matrice de sélection des composantes. Ce test est utilisé dans le cas d'une variable qualitative à plus de deux modalités.

On peut vérifier que ce test peut également s'interpréter comme un test de **corrélation partielle** entre les réponses et les variables ajoutées sachant les régresseurs du modèle sous H_0. En particulier si l'on n'a qu'un régresseur que l'on teste contre la constante seule, $p = 2, q = 1$, la statistique de Fisher ne fait intervenir que la corrélation empirique entre x et y (cf. la formule II.4).

Non-monotonicité des tests. Dans l'exemple précédent, si l'on teste H_0 : « Volume=0 » contre le modèle complet on trouve une p-value de 0,63 (c'est le résultat du test de Student déjà vu), si l'on teste H_0 : « Vitesse=0 » contre le modèle complet on trouve une p-value de 0,67 et si l'on teste H_0 : « Volume=Vitesse=0 » on trouve une p-value de 0,85. C'est-à-dire qu'on accepte plus facilement « Volume=Vitesse=0 » que « Volume=0 » ou que « Vitesse=0 » !

Ceci vient du fait qu'imposer « Vitesse=0 » (ou « Volume=0 ») change très peu RSS tandis que le changement dans le nombre de degrés de libertés va du coup favoriser l'hypothèse plus compliquée.

On rencontrera un autre exemple page 74.

III.2.4 Sélection des variables

Il s'agit de choisir les variables les plus significatives, l'idée étant d'éliminer les régresseurs dont la contribution à la prédiction, sur de nouvelles données, sera probablement nulle. Une méthode serait de tester, au vu de $\widehat{\beta}_j$ et de $\widehat{\sigma}(\widehat{\beta}_j)$ si β^*_j est significativement nul ou non, et d'éliminer le régresseur correspondant. Cette méthode ne convient cependant pas car si par exemple deux régresseurs sont très proches (la « puissance » de la « vitesse » de la page 48), un seul suffit :

$$y_i \simeq \beta_1 + \beta_2 x_2 + \beta_3 x_3 \simeq \beta_1 + (\beta_2 + \beta_3) x_2$$

mais cet algorithme éliminera probablement les deux car $\widehat{\sigma}(\widehat{\beta}_2)$ et $\widehat{\sigma}(\widehat{\beta}_3)$ sont très grands. Les hypothèses $\beta_2 = 0$ et $\beta_3 = 0$ peuvent être acceptées séparément sans que $\beta_2 = \beta_3 = 0$ le soit.

La méthode ascendante (**forward selection**) part du modèle qui ajuste y_i avec la constante seule. Pour chacun des régresseurs on calcule la valeur du RSS correspondant à son ajout au modèle et l'on choisit celui pour lequel il est le plus petit. On poursuit ainsi en ajoutant à chaque fois le régresseur minimisant le nouveau RSS. Noter que les $\widehat{\beta}_k$ changent tous à chaque fois. En présence de variables catégorielles on minimise habituellement non pas le RSS, mais un critère qui prend en compte le nombre de modalités ajoutées, souvent le critère AIC$= n \log(RSS) + 2p$, où p est le nombre de variables utilisées (on pourrait aussi prendre comme critère la statistique de Fisher associée au changement de modèle).

On s'arrête quand l'amélioration est jugée statistiquement non-significative ; ce qui se fait simplement avec le test de Fisher (équation (III.2)), ou lorsque AIC est minimal.

On met souvent cette méthode en œuvre en vérifiant après chaque étape que les régresseurs ajoutés sont encore utiles en faisant des tests de Fisher, c'est la *stepwise selection* (fonction `step()` de R) .

La méthode descendante (**backward elimination**) procède de la même façon en partant du modèle complet et en éliminant à chaque étape la variable faisant diminuer le RSS (ou autre critère en présence de variables catégorielles). On s'arrête quand la détérioration est jugée statistiquement significative par le test de Fisher correspondant à un niveau α, ou lorsque AIC est minimal

Elle peut sembler plus juste car on part d'un modèle (supposé) vrai, ce qui fait que dans les tests, au moins au début, l'hypothèse H_1 est effectivement satisfaite. En présence d'un très grand nombre de régresseurs, la méthode ascendante est souvent préférée car plus simple.

En présence d'un grand nombre de facteurs, la méthode descendante peut éliminer dès le début des variables explicatives importantes, mais perçues comme redondantes face à la masse des autres. C'est pourquoi on autorise chaque variable jugée significative à revenir dans le modèle, de manière analogue à ce qui est fait dans la méthode ascendante.

Thall, Russell et Simon [77] proposent de choisir le niveau α par validation croisée (backward elimination via repeated data splitting [BERDS]) : Séparer l'ensemble en deux, T (training) et V (validation) ; pour chaque α, faire toute la procédure backward sur T et évaluer la performance en prédiction sur V ; recommencer pour plusieurs découpages T/V ; calculer la performance moyenne pour chaque α ; faire l'estimation sur tout l'ensemble avec la valeur de α ayant conduit à la meilleure prédiction moyenne.

Autres méthodes. A ces deux dernières on préfère parfois soit le *best subset* (fonction `leaps()` de R) s'il y a peu de variables, soit le *lasso* s'il y a beaucoup de variables (`lars()`), cf. la discussion § II.9.3.

Facteurs proches, colinéarité. Il est **essentiel** de bien comprendre que si deux colonnes de X sont proches mais utiles pour la prédiction, la méthode descendante en rejettera une assez vite (car elle est redondante) pour garder l'autre longtemps. Une conclusion hâtive est de dire que le premier facteur n'est pas significatif tandis que le second l'est ; c'est évidemment inexact. Ceci vaut pour des groupes de colonnes.

Par exemple si l'on veut étudier le taux de fréquentation du médecin en fonction des deux variables « âge » et « sexe » et que les individus sont des jeunes femmes et des hommes vieux, il est clair que le plan d'expérience est mauvais, et l'on ne pourra pas démêler l'influence de l'âge de celle du sexe.

C'est pour cela qu'il est très avantageux d'avoir une matrice X la plus orthogonale (en colonnes) possible (valeurs propres de $X^T X$ presque toutes égales), c'est-à-dire un bon plan d'expérience.

Nous y reviendrons plus en détails au paragraphe Bilan 2 p. 62.

Utilisation de critères. Les méthodes présentées proposent une suite de modèles emboîtés (ou presque, en raison de du retrait possible de variables introduites), parmi lesquels on peut choisir par un critère qui s'avère être un RSS pénalisé (puisque minimiser le RSS conduirait automatiquement au modèle le plus compliqué). On a déjà vu le critère de validation croisée $\widehat{\sigma}_{CV}$, exercice 10 p. 22, et le C_p-Mallows

$$C_p = RSS + 2p\sigma_*^2$$

(où σ_* est généralement estimé sur le modèle le plus compliqué), exercice 9 p. 21. En pratique on utilise un des deux critères suivants, discutés à l'appendice A :

$$AIC = n\log(RSS) + 2p$$
$$BIC = n\log(RSS) + p\log(n)$$

où n est le nombre d'individus et p le nombre de variables. AIC est le résultat de l'estimation de l'erreur de prédiction qui serait faite sur un nouvel échantillon (c'est donc essentiellement un concurrent à $\widehat{\sigma}_{CV}$). On peut dire globalement que BIC aura tendance à choisir un modèle trop parcimonieux tandis que AIC aura la tendance inverse. Il n'est pas rare de voir AIC ajouter des variables jugées non significatives à 5% par le test de Fisher ; il donnera néanmoins un meilleur prédicteur que BIC.

Ceci permet en théorie de comparer les 2^p modèles possibles ; c'est malheureusement un principe généralement trop difficile à mettre en œuvre, pour des raisons techniques et aussi pour des raisons théoriques car essayer un trop grand nombre de modèles peut conduire encore à un phenomène de surajustement, particulièrement avec AIC.

Ces critères présentent l'avantage de permettre de comparer des modèles non emboîtés. Ils ne doivent cependant pas masquer l'importance de la statistique de Fisher pour les modèles emboîtés, qui est généralement plus précise, même si son interprétation doit être ici remise en cause car on fait ici des tests successifs, donc multiples.

Biais de choix de modèle. Dans la méthode ascendante, par exemple, le p-ième modèle aura été choisi comme le meilleur parmi toute une famille de modèles, et ce choix de régresseurs dépend des données ; cette composante aléatoire supplémentaire rend les résultats précédents en toute rigueur non-appliquables. Par exemple, si p est très grand, on se doute que $\widehat{\sigma}$ sera biaisé vers le bas.

III.2.5 Exercices

Exercice 1. Test de Chow. Le modèle est

$$y_t = a_k + b_k x_t + u_t, \quad t = 1, \dots 2T$$

avec $k = 1$ pour $t \leqslant T$, et $k = 2$ après. Cette équation modélise par exemple un changement de régime dans des données mesurées au cours du temps.

1. Mettre ce modèle sous la forme $y = X\beta + u$ pour un X bien choisi.
2. En déduire un test de H_0 :« $(a_1, b_1) = (a_2, b_2)$ » contre son contraire (on donnera L et l).

Exercice 2. On est dans le cadre linéaire gaussien habituel. Soit σ_0 une valeur nominale donnée. Proposer un test pour H_0 : « $\sigma_* = \sigma_0$ » contre H_1 : « $\sigma_* > \sigma_0$ » (on pourra, si l'on préfère, choisir d'abord la forme du test au vu des hypothèses à tester puis déterminer ensuite le seuil).

Exercice 3. On mesure le taux de cholestérol d'individus de trois groupes de taille n_1, n_2 et n_3, de sorte que l'on a en tout $n = n_1 + n_2 + n_3$ individus. Proposer une méthode pour tester si l'origine des individus (c-à-d leur groupe) influe sur leur taux de cholestérol.

***Exercice 4.** On considère le modèle

$$y = X\beta + u, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

On considère les hypothèses H_0 : « $\beta_2 = \beta_3 = 0$ » et H_1 : « $\beta_3 = 0$ ». On note RSS_0, RSS_1 et RSS_2 les résidus calculés respectivement sous les hypothèses H_0, H_1 et le modèle complet, et $\widehat{\beta}$ et $\widehat{\sigma}$ les estimées sous le modèle complet. Trouver la loi sous H_0 de la statistique

$$\frac{(RSS_0 - RSS_1)/q}{RSS_2/(n-p)}$$

où p est la dimension de β et q celle de β_2. On commencera par remarquer que $RSS_0 - RSS_1$ est fonction de $\widehat{\beta}$ (cf. lemme 19), et que RSS_2 est fonction de $\widehat{\sigma}$.

En déduire un test de H_0 contre H_1.

Exercice 5. Une estimation OLS à 40 individus donne les résultats suivants (β_1 est le coefficient du régresseur constant) :

$$\begin{aligned} \widehat{\beta} &= (32 \quad 8 \quad -4 \quad -1)^T \\ RSS &= 18 \\ TSS &= 80 \\ (X^T X)^{-1} &= \begin{pmatrix} 20 & \times & \times & \times \\ \times & 1 & 1 & \times \\ \times & \times & 2 & 1 \\ \times & \times & \times & 4 \end{pmatrix}. \end{aligned}$$

Tous les test seront faits avec un seuil de confiance de 95%. On pourra utiliser les valeurs suivantes des quantiles de la loi de Fisher-Snedecor

$$f_{2,36}(0,95) \simeq 3,27, \quad f_{3,36}(0,95) \simeq 2,87, \quad f_{4,32}(0,95) \simeq 2,67$$

et ceux de la loi de Student

$$t_{36}(0,975) \simeq 2,03, \quad t_{36}(0,95) \simeq 1,69.$$

1. Calculer un estimateur sans biais de σ_*^2.
2. Faire pour chaque régresseur le test de nullité du coefficient.

3. Faire le test de nullité simultanée de tous les coefficients, sauf β_1.
4. Faire le test H_0 : « $\beta_2 + \beta_3 = 7$ » contre son contraire.
   On basera le test sur la statistique $S = (7 - \widehat{\beta}_2 - \widehat{\beta}_3)/\widehat{\tau}$, $\widehat{\tau} = \widehat{\sigma}(7 - \widehat{\beta}_2 - \widehat{\beta}_3)$ dont on donnera la loi sous H_0 (cf la statistique T_u de la proposition 18).
5. Faire le test H_0 : « $\beta_2 + \beta_3 = 7$ » contre H_1 : « $\beta_2 + \beta_3 < 7$ ». On utilisera S. Qu'observe-t-on ?
6. Faire le test H_0 : « $\beta_3 = \beta_4 = 0$ » contre son contraire.
7. Sur les 20 premières observations on a obtenu

   $$\widehat{y}_i = 35 + 6x_{i1} - 2x_{i2} - 2x_{i3}, \quad RSS = 7$$

   et sur les 20 dernières

   $$\widehat{y}_i = 29 + 9x_{i1} - 5x_{i2} - 3x_{i3}, \quad RSS = 6.$$

   Peut-on considérer que β n'a pas changé ? On pourra s'inspirer de l'idée développée dans l'exercice 1 p. 52.

Exercice 6. (IC en prédiction). On considère le modèle habituel sur lequel on a obtenu une estimée $\widehat{\beta}$ de β^*. On cherche un intervalle de confiance pour $x'\beta^*$ et y' où (x', y') est une paire régresseur/réponse satisfaisant le modèle.

1. Soit $h = x'(X^TX)^{-1}x'^T$. Quelle est la loi de $(x'\widehat{\beta} - x'\beta^*)/\widehat{\sigma}h^{1/2}$?
   En déduire un intervalle de confiance $I(x)$ centré en $x'\widehat{\beta}$ et de niveau α pour $x'\beta^*$.
2. Quelle est la loi de $(x'\widehat{\beta} - y')/\widehat{\sigma}\sqrt{1+h}$?
   En déduire un intervalle de confiance centré en $x'\widehat{\beta}$ et de niveau α pour y'.

***Exercice 7. (IC simultanés en prédiction).** L'intervalle de confiance de l'exercice précédent satisfait $P(x'\beta^* \in I(x')) \geqslant 1 - \alpha$, mais si l'on veut des prédicteurs pour plusieurs régresseurs simultanément, par exemple x' et x'', et sans faire baisser le niveau, il faudrait pouvoir assurer

$$P(x'\beta^* \in I(x') \quad \text{et} \quad x''\beta^* \in I(x'')) \geqslant 1 - \alpha$$

qui n'est pas satisfait. La suite de l'exercice propose une solution à ce problème.

1. Soit Q une matrice carrée telle que $QQ^T = (X^TX)^{-1}$. Montrer qu'il existe une variable normale standard Γ indépendante de $\widehat{\sigma}$ telle que $\widehat{\beta} - \beta^* = \sigma_* Q\Gamma$.
2. Montrer que $\|xQ\|^{-2}(x\widehat{\beta} - x\beta^*)^2$ est majoré à un facteur près par un χ^2_p indépendant de x.
3. Montrer que

   $$J(x) = [x\widehat{\beta} - \delta(x), x\widehat{\beta} + \delta(x)], \quad \delta^2 = p\widehat{\sigma}^2(x(X^TX)^{-1}x^T)f_{p,n-p}(1-\alpha)$$

   est un intervalle de confiance uniforme en x de niveau α, c-à-d que pour toute valeur de β^*

   $$P(\forall x, \quad x\beta^* \in J(x)) \geqslant 1 - \alpha.$$

Exercice 8. (**Lien avec le rapport de vraisemblance**) Vérifier que la statistique du test de Fisher vaut

$$\frac{n-p}{p-p_0}(\lambda^{2/n} - 1) \tag{III.3}$$

où λ est le rapport de vraisemblance $P(y)/P_0(y)$, calculé avec les estimateurs au maximum de vraisemblance. On pourra utiliser l'exercice 1 page 45.

On pourra comparer cette statistique avec celles présentées à l'annexe C (faire n grand).

III.3 Analyse des résidus. Mesures d'influence

L'approche la plus simple est le tracé de l'histogramme des résidus, qui permet de confirmer l'hypothèse gaussienne et également de détecter des individus qui ne suivent pas le modèle (résidus anormalement grands). On peut toutefois faire une étude plus précise. Rappelons que la loi de $\widehat{u}_i$ est $\mathcal{N}(0,(1-h_i)\sigma_*^2)$ (car $\widehat{u}=Ku$, cf. l'exercice 9 p. 21).

20 - Définition

On appelle résidus studentisés les estimateurs centrés « réduits » des erreurs

$$r_i = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}\sqrt{1-h_i}}$$

On appelle résidus studentisés par validation croisée

$$r_i^* = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_i}} = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}. \tag{III.4}$$

Les notations sont celles du § II.2.5, et la dernière égalité est laissée en exercice.

21 - Proposition

Si $y \sim \mathcal{N}(X\beta^*, \sigma_* I)$, r_i^* suit une loi de Student de paramètre $n-p-1$.

La démonstration est immédiate au vu des résultats du paragraphe III.2 et de la formule $\widehat{u}_i = (1-h_i)(y_i - x\widehat{\beta}_{(i)})$ conséquence de (II.1), qui assure l'indépendance des deux termes de la fraction. La statistique r_i^* est en pratique préférée à r_i.

22 - Définition

Une **donnée aberrante** au niveau α est un individu i pour lequel r_i^* dépasse le seuil donné par la loi de Student pour un risque d'erreur α.

Un niveau α raisonnable est $1/n$, soit un seuil à $t_{n-p-1}(1-1/(2n))$. Si n est grand, on peut être tenté de choisir α plus grand, par exemple $0,05$ mais ce que l'on détecte alors devrait plutôt être appelé des individus extrêmes, puisque statistiquement de tels individus seront toujours présents en proportion de 5%. Idéalement, dans une approche théorique, les données aberrantes sont celles qui ne sont pas conformes au modèle, ou erronées ; la définition ci-dessus est utilisée à des fins pratiques, avec les limitations que l'on vient de voir. Pour la motivation de la définition suivante, on réfère au § II.2.5 p. 15 :

23 - Définition

Une **donnée isolée** au niveau α est un individu i pour lequel nh_i/p dépasse le seuil α (souvent choisi à 3). Une **donnée atypique** est une donnée soit isolée soit aberrante.

Interprétation. Une donnée aberrante est donc une donnée dont la réponse y est peu conforme au modèle estimé sur les autres données. Une donnée isolée est une donnée dont le régresseur x est isolé dans l'espace ; son retrait augmenterait donc sensiblement la matrice de covariance de $\widehat{\beta}$; de plus $\widehat{\beta}$ est très sensible à la valeur de y correspondante (**effet levier**). Les **données influentes** sont celles qui influent sur l'estimation de β^*. La mesure d'influence la plus utilisée est la distance de Cook qui vaut

$$C_i = \frac{(\widehat{\beta}-\widehat{\beta}_{(i)})^T X^T X(\widehat{\beta}-\widehat{\beta}_{(i)})}{\widehat{\sigma}^2 p} = \frac{h_i}{p(1-h_i)} r_i^2 \tag{III.5}$$

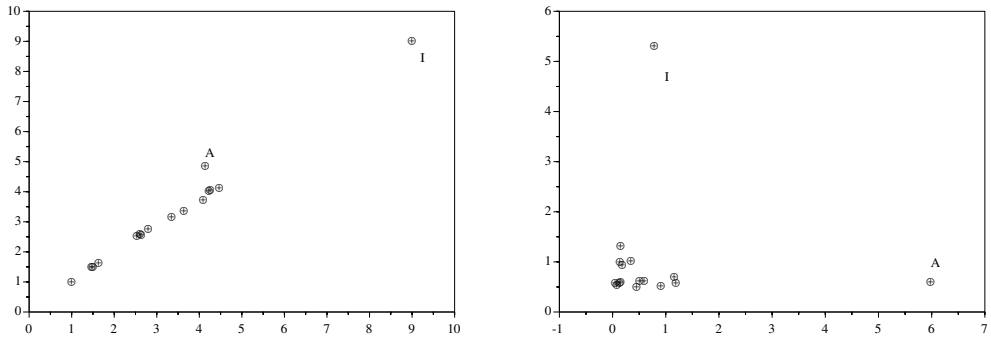


FIGURE III.1 – Sur la première figure x_i est en abscisse et y_i est en ordonnée. Le point A est aberrant et I est isolé. La seconde figure représente les nh_i/p en fonction des résidus studentisés (valeur absolue).

(la dernière identité est laissée en exercice). Comme une grande valeur de cette statistique peut aussi bien venir d'un fort résidu que de l'isolement dans l'espace de l'individu, le meilleur tracé est a priori celui des $(r_i^*, nh_i/p)$, comme figure III.1.

Bilan. Même si le modèle gaussien semble douteux, les résidus studentisés sont intéressants car ils sont normalisés, et l'on peut faire le graphique de la figure III.1, quitte à rester prudent dans l'interprétation qu'on lui donne.

Il faut se garder d'éliminer de but en blanc des individus aberrants d'une analyse pour la recommencer ensuite ; un tiers pourrait y voir à juste titre une manipulation grossière pour biaiser l'étude. L'analyse des résidus se contente de jeter la suspicion sur certains individus et c'est ensuite à l'analyste d'essayer de savoir s'ils contiennent des erreurs (de mesure, etc.), et si ce n'est pas le cas, ils peuvent être au contraire importants pour l'estimation ou la remise en cause du modèle (linéarité, etc.).

C'est une grave erreur que d'éliminer a priori les individus isolés, qui au contraire peuvent être porteurs de beaucoup d'information. On peut cependant être amené à le faire pour améliorer la linéarité du modèle, considérant que le modèle linéaire n'est généralement qu'une approximation raisonnable, valide sur un domaine restreint.

Pour les exercices suivants, il pourra être utile de se servir du théorème 9 p. 16.

Exercice 1. Pourquoi r_i ne suit-il pas une loi de Student ?

Exercice 2. Montrer que $C_i = \dfrac{|x_i(\widehat{\beta}_{(i)} - \widehat{\beta})|^2}{ph_i\widehat{\sigma}^2}$.

Exercice 3. Démontrer la deuxième égalité de (III.4). Démontrer la deuxième égalité de (III.5).

III.4 Analyse de la variance. Aspects pratiques

III.4.1 Analyse de la variance à un facteur

L'analyse de variance proprement dite s'intéresse à la situation où les variables explicatives sont catégorielles. Dans ce paragraphe il n'y en a qu'une.

On dispose de $n = n_1 + .. + n_p$ observations y_{ik}, $i = 1, \dots p$, $k = 1, \dots n_i$. i est l'indice de groupe et n_i la taille du groupe i. Par exemple, supposons que l'on veuille tester p différents engrais ; pour chaque engrais i, on fera n_i expériences (plantations) et y_{ik} désignera la production du k-ième champ test utilisé pour le i-ième type d'engrais. Le modèle de régression correspondant est

$$y_{ik} = \mu_i + u_{ik} \tag{III.6}$$

où μ_i est la productivité du i-ième engrais. On met cette régression sous la forme $y = X\beta + u$:

$$y = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{1} \end{pmatrix} \mu + u$$

où $\mathbf{0}$ et $\mathbf{1}$ sont des vecteurs de 0 et de 1. Le vecteur ligne x_i indique quel engrais a été utilisé. La matrice $X^T X$ est la diagonale des n_i et

$$\widehat{\mu}_i = \bar{y}_{i\cdot} = n_i^{-1} \sum_k y_{ik}.$$

On s'intéresse à savoir si les $\widehat{\mu}_i$ sont significativement différents, c'est-à-dire à tester H_0 : « $\mu_1 = \dots = \mu_p$ », c'est-à-dire si l'engrais a un effet mesurable. Les quantités intervenant dans le test de Fisher sont

$$\sum_i (y_i - \bar{y})^2 = \sum_i n_i (\bar{y}_{i\cdot} - \bar{y})^2 + \sum_{ik} (y_{ik} - \bar{y}_{i\cdot})^2$$
$$\text{TSS} \quad = \quad \text{ESS} \quad + \quad \text{RSS}$$

avec ici $RSS_0 = TSS$. Chacune des trois statistiques est un χ^2 sous H_0, cf. § III.2.3.

Il faut bien voir que ESS et RSS s'interprètent comme les variances interclasse et intraclasse, et la statistique de Fisher est proportionnelle à leur rapport.

Paramétrisation habitelle. Le même modèle peut se réécrire

$$y_{ik} = \mu + \alpha_i + u_{ik}, \quad \alpha_I = 0.$$

Le nombre de paramètres libres est toujours p. Cette paramétrisation est celle généralement employée par les logiciels. Le coefficient μ est qualifié d'« intercept ». La modalité J est appelée « modalité de référence » car $\mu = \mu_I$, et $\alpha_i = \mu_i - \mu$ mesure l'écart à cette dernière.

Exemple. On s'intéresse à la composition des hotdogs [1]. La réponse est la teneur en calories et la variable explicative Viande a trois modalités : Volaille, Bœuf, Divers (essentiellement porc et bœuf). Il y a 54 individus. La table d'analyse de variance du modèle (commande ANOVA de R) montre bien que le type de viande influe significativement sur les calories, expliquant plus d'un tiers (39%) de la variabilité des données :

	dl	Sum Sq	Pr(>F)
Viande	2	17700	3,8e-06
Residuals	51	28000	

L'analyse des coefficients du modèle (SUMMARY de R) donne

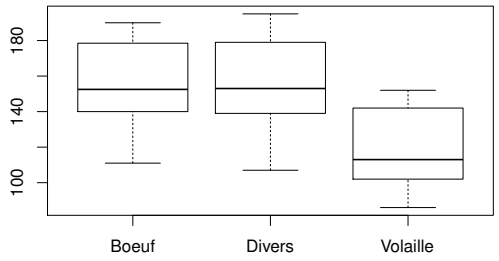
	Estimate	Std. Error	Pr(>|t|)
(Intercept)	157	5,2	< 2e-16
Divers	2	7,7	0,8
Volaille	- 38	7,7	9,4e-06

Attention, l'interprétation du 0,8 est que les viandes diverses n'ont pas d'apport calorique significativement différent du bœuf (associé lui-même à un coefficient nul : l'analyse est dissymétrique). Si l'on s'arrange pour que ce soit la variable Volaille qui ait son coefficient nul, on obtient la table suivante

1. The Data and Story Library. `lib.stat.cmu.edu/DASL`. Hot dogs story.

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	119	5,7	< 2e-16
Boeuf	38	7,7	9e-6
Divers	40	8	8e-06

où l'on voit que les coefficients de Bœuf et Divers sont significativement différents de 0 (donc de Volaille) mais sans doute indistinguables entre eux vu l'écart-type. Cette analyse se confirme par une représentation en boîtes à moustaches :



Si l'on regroupe ces deux classes on obtient après une analyse supplémentaire le modèle :

$$\text{Calories}=157,7-37\ 1_{\text{Volaille}}+bruit, \qquad \widehat{\sigma}=23, \qquad R^2=0,39.$$

Ce regroupement peut se justifier plus précisément en testant le modèle aggrégé contre le modèle original (commande ANOVA de R avec deux arguments).

III.4.2 Analyse de la variance à deux facteurs

Supposons que l'on veuille maintenant tester différents engrais dans différentes régions et voir si certains engrais sont plus adaptés à certaines régions. Les observations seront maintenant y_{ijk} où $i = 1, \dots I$ est l'indice d'engrais (premier facteur, noté A dans la suite), $j = 1, \dots J$ est l'indice de région (facteur B), et k l'indice d'expérience, qui varie entre 1 et n_{ij}. Le plan d'expérience est dit complet si tous les n_{ij} sont strictement positifs, et équilibré s'il sont égaux. On notera $n_{i.} = \sum_j n_{ij}$ et de même pour $n_{.j}$. On supposera d'abord que $n_{ij} > 0$ pour tous i, j.

Modèle complet avec interactions. C'est le modèle :

$$AB: \quad y_{ijk} = \mu_{ij} + u_{ijk} \tag{III.7}$$

Modèle additif. C'est le modèle pour lequel μ_{ij} est somme de deux termes, $\mu_{ij} = \mu_i + \nu_j$:

$$A + B: \quad y_{ijk} = \mu_i + \nu_j + u_{ijk}. \tag{III.8}$$

C'est un modèle à $I + J - 1$ paramètres libres car les μ_i et ν_j ne sont définis qu'à une constante près. Pour définir les paramètres de manière unique, on impose une contrainte, par exemple $\nu_J = 0$.

Ce modèle correspond au précédent sous les contraintes $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$.

Modèles à un facteur. Ce sont les modèles :

$$A: \quad y_{ijk} = \mu_i + u_{ijk}$$
$$B: \quad y_{ijk} = \mu_j + u_{ijk}.$$

Ceci correspond aux contraintes $\mu_{ij} = \mu_{ij'}$ ($\mu_{ij} = \mu_{i'j}$ pour le second).

Paramétrisation habituelle. On préfère souvent utiliser une autre paramétrisation, qui met mieux en valeur la contribution des différents facteurs et de l'interaction :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{III.9}$$

soit le modèle

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + u_{ijk}. \tag{III.10}$$

Le terme γ_{ij} est nul dans le cas du modèle additif. Pour avoir unicité de la décomposition, les logiciels habituels utilisent les contraintes :

$$\alpha_I = \beta_J = \gamma_{Ij} = \gamma_{iJ} = 0, \quad i = 1, \dots I, \quad j = 1, \dots J. \tag{III.11}$$

D'où la table de décompte des paramètres libres

Facteur	μ	$\alpha_.$	$\beta_.$	$\gamma_{..}$	total
Param. libres	1	$I-1$	$J-1$	$(I-1)(J-1)$	IJ

Exemple. [40] Il s'agit d'étudier l'effet de l'âge sur la mémoire. 5 types d'expérience sont faites au terme desquelles on demande toujours au sujet de se remémorer certains mots. La réponse est le nombre de mots cités (on verra plus loin qu'un modèle linéaire généralisé serait plus adapté à ce type de réponse). Le type d'expérience est la variable « Process », de modalités « Counting », « Imagery », « Intentional », « Rhyming » et « Adjective ». La variable « Age » est ajoutée. Il y a 100 individus, correspondant à 10 répétitions des 10 combinaisons possibles des modalités. Voici la table d'analyse des coefficients (procédure SUMMARY de R, option `/solution` dans la procédure GLM de SAS) :

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	11	0.9	< 2e-16 ***
AgeYounger	3.8	1.27	0.0035 **
ProcessCounting	-4	1.27	0.002 **
ProcessImagery	2.4	1.27	0.06 .
ProcessIntentional	1	1.27	0.43
ProcessRhyming	-4.1	1.27	0.0017 **
AgeYounger :ProcessCounting	-4.3	1.8	0.018 *
AgeYounger :ProcessImagery	0.4	1.8	0.824
AgeYounger :ProcessIntentional	3.5	1.8	0.054 .
AgeYounger :ProcessRhyming	-3.1	1.8	0.087 .

$R^2 = 0.7$

$\widehat{\sigma} = 2.8$

Cette table est difficile à lire mais on voit un effet indiscutable de l'âge, des résultats significativement différents d'une expérience à l'autre, et une similarité possible entre les l'effet des modalités Counting et Rhyming. La table de statistique de Fisher (procédure ANOVA de R) concerne l'effet global des variables, elle est plus facile à lire, son interprétation exacte en termes de tests sera précisée plus bas :

	Df	Sum Sq	Pr(>F)
Age	1	240	4e-07 ***
Process	4	1514	< 2.2e-16 ***
Age :Process	4	190	0.00028 ***
Residuals	90	722	

Attention PRUDENCE. Il faut bien voir que les μ, α_i, β_j et les γ_{ij} DÉPENDENT DES MODALITÉS DE RÉFÉRENCE CHOISIES, ET N'ONT PAS DE SIGNIFICATION PRIS SÉPARÉMENT, ce qui fait que ces paramètres présentent peu d'intérêt pour l'utilisateur. On se gardera donc bien de les interpréter hâtivement. Par exemple dans un modèle avec interaction, $\widehat{\alpha} = 0$ ne signifie rien de particulier.

Considérons par exemple le modèle additif ; le test « $\alpha_i = 0$ » dépend de la convention, par exemple dans le cas de (III.11) il signifie en réalité que i et J ont même effet. C'est pourquoi les logiciel refuserons toujours de faire ce genre de test (c'est la théorie des « testable functions »). En revanche le logiciel acceptera de tester « $\alpha_i - \alpha_j = 0$ », qui a le même sens *indépendamment de la convention utilisée* et qui s'interprète comme « i et j ont même effet ». Dans le modèle avec interactions, la situation est encore

plus confuse. En pratique, si l'on veut confirmer ou infirmer que « i et j ont même effet » (come le Bœuf et Divers dans la table p. 57, ou Counting et Rhyming dans la précédente), il pourra être plus simple de tester le modèle où les deux classes ont été fusionnées contre l'original.

Cas du plan incomplet. Si n_{ij} n'est pas toujours positif, mais que $\sum_j n_{ij}$ et $\sum_i n_{ij}$ sont non-nuls, le nombre de paramètres du modèle complet n'est plus IJ mais le nombre de n_{ij} non-nuls : c'est le nombre de μ_{ij} en jeu, les cellules vides étant considérées comme inexistantes. Pour calculer les termes p et p_0 intervenant dans la statistique de Fisher, une méthode qui marche toujours (indépendamment du nombre de facteurs) de prendre le rang des matrices X correspondant aux deux modèles en compétition : $p = r$, $p_0 = r_0$.

Plan d'expérience équilibré et sommes de carrés. Cette situation a l'avantage de faciliter l'interprétation des analyses car il n'y a pas de facteurs proches. De plus elle présente des formules simples pour les estimées.

Soit K la valeur commune des n_{ij}. Alors $n = IJK$. On vérifie que les estimées pour le modèle complet (III.10) sous les contraintes $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ sont

$$\widehat{\mu} = \bar{y}$$
$$\widehat{\alpha}_i = \bar{y}_{i..} - \bar{y} = \frac{1}{JK}\sum_{jk} y_{ijk} - \bar{y}, \qquad \widehat{\beta}_j = \bar{y}_{.j.} - \bar{y} = \frac{1}{IK}\sum_{ik} y_{ijk} - \bar{y}$$
$$\widehat{\gamma}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}$$

et que ces estimées valent aussi pour les modèles additifs ($\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j$) et « à un facteur » ($\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i$, et $\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\beta}_j$). Les variances expliquées par chaque facteur sont définies et calculées comme suit (RSS_M désigne le RSS du modèle M) :

$$SS_\alpha = TSS - RSS_A = JK\sum_i \widehat{\alpha}_i^2, \qquad SS_\beta = TSS - RSS_B = IK\sum_j \widehat{\beta}_j^2$$
$$SS_\gamma = RSS_{A+B} - RSS_{AB} = K\sum_{ij} \widehat{\gamma}_{ij}^2$$

(vérification laissée en exercice). Notons que la décomposition $y_{ijk} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij} + \widehat{u}_{ijk}$ est orthogonale (5 vecteurs orthogonaux de dimension IJK) en raison de la paramétrisation et de l'équilibre, ce qui revient à dire qu'on a décomposé l'espace $\mathcal{X}$ en quatre sous-espaces orthogonaux définis par les contraintes imposées, correspondant chacun au facteur, « constante », A, B ou « interaction » ; SS_x est le carré de la norme de la projection de y sur le sous-espace correspondant et le théorème de Pythagore implique la formule de sommation :

$$\begin{array}{lccccccccc} SS & : & TSS & = & SS_\alpha & + & SS_\beta & + & SS_\gamma & + & RSS_{AB} \\ d.l. & : & n-1 & = & I-1 & + & J-1 & + & (I-1)(J-1) & + & n-IJ \end{array} \qquad \text{(III.12)}$$

où d.l. est le nombre de degrés de liberté de SS_x sous $x = 0$. La variance totale est décomposée en un terme de bruit RSS et trois termes distincts dûs respectivement à la variabilité en fonction de la première variable, de la seconde, et de l'interaction des deux. Chaque terme est interprété comme la contribution de chaque facteur (cf. l'explication de la page 15 pour la justification de l'usage des carrés). C'est la colonne « Sum Sq » de la table de la page 58. Sous l'hypothèse nulle que $\alpha_\cdot = \beta_\cdot = \gamma_{\cdot\cdot} = 0$, ces quatre termes sont, après division par σ_*^2, des χ^2 indépendants avec les degrés de liberté correspondants. Ces termes sont ceux qui servent à construire les statistiques de Fisher. La simplicité de cette décomposition et de son interprétation sont caractéristiques du plan équilibré, et des plans où X est orthogonale (ici, les espaces associées à A, B et $A.B$ au travers de la paramétrisation choisie sont orthogonaux).

Plan d'expérience équilibré incomplet. Soit trois facteurs à deux modalités et les 4 expériences suivantes (au lieu de 8 pour avoir un plan complet) répétées K fois chacune :

	A	B	C
Exp1	0	0	0
Exp2	0	1	1
Exp3	1	0	1
Exp4	1	1	0

Ici $n = 4K$ et la matrice X est le tableau ci-dessus où les lignes sont répétées K fois et une colonne de 1 ajoutée. L'équilibre se voit à la symétrie du plan par rapport aux facteurs qui fait que les colonnes recentrées sont orthogonales. On a encore la décomposition en sommes de carrés

$$TSS = SS_A + SS_B + SS_C + RSS_{A+B+C}$$

où chaque terme est interprété comme la contribution de chaque facteur.

Modèle de base additif. Dans l'exemple précédent seules 4 combinaisons de facteurs ont été considérées sur les 8 possibles, on ne peut donc pas estimer plus de 4 paramètres. On part alors du modèle additif.

De même, si dans un modèle à deux facteurs $n_{ij} = 1$ pour tous i et j, le modèle (III.7) ne présente pas d'intérêt et les tests seront impossibles car RSS$= 0$ et $n = IJ$.

On peut aussi faire ce choix si le test « $\gamma = 0$ » donne une réponse que l'on considère définitive.

III.4.3 Interprétation des tables

Nous avons discuté, p.ex. au § III.4.1, des tables d'analyse des coefficients. Nous nous intéressons ici aux tables qui étudient les facteurs globalement.

Les tests de modèles sont souvent interprétés comme des tests de significativité des différents facteurs. On les illustre ici dans le cas de trois facteurs qui peuvent être soit qualitatifs soit quantitatifs. La table III.2 montre les deux types de tests commentés plus bas.

(I)

Facteur à tester	H_1	H_0
A	A	cst
B	$A+B$	A
interaction	AB	$A+B$
C	$AB+C$	AB

(III)

Facteur à tester	H_1	H_0
A	$AB+C$	$B+C+A.B$
B	$AB+C$	$A+C+A.B$
C	$AB+C$	AB
interaction	$AB+C$	$A+B+C$

TABLE III.2 – Tests de type I et de type III pour le modèle $AB+C$. (À gauche) Tests emboîtés : ce qu'ajoute chaque facteur aux précédents. (À droite) Test contre $H_1 =$« complet » : ce qu'ajoute un facteur quand les autres sont présents. La partie $A.B$ désigne les termes d'interaction seuls

La notation $A.B$ correspond au facteur obtenu par orthogonalisation des colonnes de AB à celles correspondant à A et à B ; ce sont des termes d'interaction pure d'interprétation obscure, qui sont les $\gamma_{..}$ du § III.4.2 pour une paramétrisation particulière.

► **Test d'hypothèses emboîtées :** Il a le mérite considérable de la simplicité d'interprétation. Le statisticien définit une suite croissante de modèles par ajout des facteurs un à un. On teste chaque modèle contre le précédent. Typiquement la méthode descendante dictera l'ordre à choisir (cf. § III.2.4).

Le test dit « de type I » (procédure ANOVA de R ou S+) est emboîté, mais le dénominateur de la statistique de Fisher est en fait remplacé par le RSS du modèle complet avec la modification correspondante du seuil, cf. l'exercice 4 p. 52. En conséquence, pour un plan équilibré, l'ordre d'introduction des facteurs n'intervient pas, cf. l'exercice 10 p. 68.

Dans le cas de facteurs proches le premier sera choisi et l'autre rejeté, cf. § III.2.4.

EXEMPLE : Reprenons l'exemple de la prédiction de la consommation en fonction du volume, du poids, de la puissance et de la vitesse maximale (table III.1 page 48). On obtient les analyses de type I suivantes

Type I	Pr
Volume	< 0.0001
P2	< 0.0001
Puissance	<0.0001
Vitesse	0.7

Type I	Pr
Volume	< 0.0001
P2	< 0.0001
Vitesse	<0.0001
Puissance	0.83

Type I	Pr
P2	< 0.0001
Vitesse	<0.0001
Volume	0.6
Puissance	0.83

On voit bien sur les deux premiers tableaux l'effet de la colinéarité de la vitesse et de la puissance. Le troisième indique que le volume n'ajoute rien au poids et la vitesse. Pour les données du § II.5.1 on obtient les résultats suivants

Type I	Pr
x	0.084
x^2	0.0014
x^3	0.0064
x^4	0.68

et les résultats sont similaires pour des degrés plus élevés. Le test de type I donne de bons résultats car ici la situation est très différente. Le plan d'expérience est encore très déséquilibré, mais le fait que le test de type I dépende de l'ordre des facteurs n'est pas ici un inconvénient car il y a un ordre des facteurs bien déterminé, et donc une suite croissante de modèle clairement définie.

▶ **Tests contre H_1 =« complet ».** C'est le test de H_1 contre H_0 du tableau III.2 (Type III de SAS. Avec R utiliser la commande `Anova(lm( ...),type='III')` de la bibliothèque CAR (ou `summary` en absence de variable catégorielle)). Il est fortement remis en question par la communauté scientifique [22], en raison de la présence d'interaction sans les facteurs principaux dans H_0. Ce type ne conduit pas à une décomposition exacte de la variance en somme de carrés. En absence d'interaction, il correspond au test de nullité de la page 48, transformé en test de Fisher pour les facteurs à plus de deux modalités.

Dans le cas d'interactions, le test de AB contre $B + A.B$ pour le facteur A (on oublie C pour simplifier) est une extension au cas déséquilibré du test de $\alpha = 0$ dans le plan équilibré de la page 59. On teste donc les facteurs simples contre le modèle complet en gardant les interactions, ce qui est très discutable car on peut difficilement imaginer une interaction AB sans que A soit significatif ! D'un point de vue assez approximatif, A sera rejeté si son effet est totalement imprévisible si l'on ne connaît pas B [2].

Les résultats ne dépendent pas de l'ordre dans lequel sont présentés les facteurs.

La structure du test fait qu'il a clairement **tendance à rejeter les facteurs proches**.

Il faut donc l'utiliser pour illustrer la contribution **additionnelle** de chaque facteur et leur **significativité**. Ses conclusions de significativité sont fiables : dans les exemples précédents l'analyse élimine les facteurs :

Type III	Pr
Volume	0.63
P2	< 0.0001
Puissance	0.83
Vitesse	0.7

Type III	Pr
x	0.35
x^2	0.32
x^3	0.45
x^4	0.68

2. Dans le test de significativité de A de SAS, l'hypothèse H_0 ($B + C + A.B$ dans le tableau) consiste à supposer que l'effet moyen de A au sens où l'on fait la moyenne des effets quand B varie, est indépendant de la valeur choisie pour A. Sur le modèle suivant où A possède 3 modalités et B en possède 2 (on oublie C pour simplifier)

$$y = \beta_1 + \beta_2 1_{A=1} + \beta_3 1_{A=2} + \beta_4 1_{B=1} + \beta_5 1_{A=1,B=1} + \beta_6 1_{A=2,B=1} + u$$

l'hypothèse pour tester l'influence de A sera

$$\begin{cases} 2\beta_2 + \beta_5 = 0 \\ 2\beta_3 + \beta_6 = 0 \end{cases}$$

De même celle pour tester B : $3\beta_4 + \beta_5 + \beta_6 = 0$.

► **La présentation habituelle** de ces résultats d'analyse de variance consiste en général à donner sur chaque ligne du tableau le facteur à tester puis le « $RSS_0 - RSS$ » correspondant au test, les degrés de liberté, la statistique de Fisher, et enfin le niveau de signification obtenu. La somme des SS ainsi présentés fait, au moins dans le cas du test de type I, le TSS, illustrant la contribution de chaque facteur à la variance totale (cf. la colonne SS du § III.4.4, et l'exercice 10 p. 68).

► **Type I et Type III.** Le but originel des tests est de montrer la significativité de certains facteurs ; dans cette optique il convient de s'arranger à l'avance pour avoir un plan d'expérience correct, et le mieux est de tester entre des hypothèses claires ; comme alternative, le test de type III peut se justifier (car le plus sévère et donc le plus convaincant), bien qu'il soit très remis en cause [22]. En revanche, l'utilisation des tests pour faire de la sélection de modèle ne doit être vue que comme une application supplémentaire, avec une mise en pratique assez informelle, mais qui réclame une bonne compréhension de la situation ; ici les tests de type I sont pratiques et permettent d'illustrer graphiquement par les différents SS la contribution de chaque facteur, avec prudence car l'ordre de leur introduction importe.

► **Bilan 1 : facteurs simples et interactions.** Si l'interaction AB est considérée comme significative alors A et B le sont (les deux premières lignes du tableau III.2 ne sont plus des tests de significativité). Dans le cas contraire on peut préférer l'éliminer du modèle et reprendre l'analyse ; on peut également tester A directement par le modèle complet contre le modèle sans A ($B + C$ contre $AB + C$).

► **Bilan 2 : facteurs proches, colinéarité.** La difficulté vient des facteurs significatifs proches ; pour les faire apparaître dans les tests, l'idée est que l'élimination de l'un rendra l'autre significatif dans le modèle ; on peut par exemple comparer les résultats de différents tests simples (p.ex. modèles additifs faisant intervenir un des facteurs, l'autre ou les deux, ou bien tests emboîtés en changeant l'ordre des facteurs (on peut les faire sous R avec ANOVA(H_1,H_0))). Une ACP des régresseurs ou une analyse des correspondances confirmera les doutes.

En règle générale, s'il n'y a pas de facteurs proches, les conclusions seront faciles à tirer. S'il y en a, il faut analyser cette proximité ; elle peut avoir deux origines :

- Corrélation effective (dans le monde réel) entre différentes variables (on postule donc l'existence d'une distribution pour les régresseurs) ; par exemple la cylindrée et la puissance.
- Plan d'expérience déséquilibré : les sujets âgés sont massivement des femmes. C'est une corrélation « artificielle » (entre sexe et âge) introduite par le choix des individus.

DANS LE PREMIER CAS la conclusion est simple puisqu'en gros *le facteur éliminé a une influence au travers de sa corrélation avec le facteur conservé.*

DANS LE DEUXIÈME CAS il est difficile de conclure puisque les données sont intrinsèquement mauvaises : si l'on veut étudier le taux de fréquentation du médecin en fonction de l'âge et du sexe et que les individus de l'échantillon sont des femmes âgées et des hommes jeunes, il est clair que le plan d'expérience est mauvais, et l'on ne pourra pas démêler l'influence de l'âge de celle du sexe ; on conclura alors à *l'effet globalement significatif des deux facteurs sans pouvoir distinguer lequel importe vraiment.*

► **Variable significative décorrélée.** Il peut arriver qu'une des variables explicatives ait une corrélation quasi-nulle avec les réponses (et soit même rejetée dans l'analyse de type I) mais qu'elle soit significative dans l'analyse de type III : ceci vient du fait qu'elle est implicitement présente dans d'autres régresseurs importants, ce qui peut rendre sa corrélation partielle avec y importante.

Considérons par exemple les variables «Criminalité» (réponse), «Revenu médian» et «Inégalité» (pourcentage de famille en dessous de la moité du revenu médian) des données d'Ehrlich[3]. On trouve en sortie d'analyse du modèle Crime=Rev+Ineq (commande `summary`) les résultats suivants :

Type III	Pr
Rev	0.000476
Ineq	0.000476

Type III	Pr
Ineq	0.229

Type III	Pr
Revorth	5.32e-06
Ineq	0.132836

Le deuxième tableau indique une corrélation non significative entre «Crime» et «Ineq» ; cependant le

3. Disponible sur OzDASL, `http://www.statsci.org/data/general/uscrime.html`

premier tableau signale que cette variable importe. La variable «Revorth» du troisième est l'erreur de prédiction de «Rev» par «Ineq» : Revorth=Rev- b.Ineq, b=cov(Rev,Ineq)/var(Ineq). On voit que finalement seul «Revorth», le supplément d'information que «Rev» apporte sur «Ineq», importe.

III.4.4 Un exemple à trois facteurs

Des étudiants essayent des avions en papier avec deux types de pliage (facteur T), deux sortes de papier (facteur P) et deux types d'angle de lancer (facteur A) [59]. La réponse est la distance D parcourue. Il y a deux individus pour chaque combinaison de facteurs soit 16 en tout.

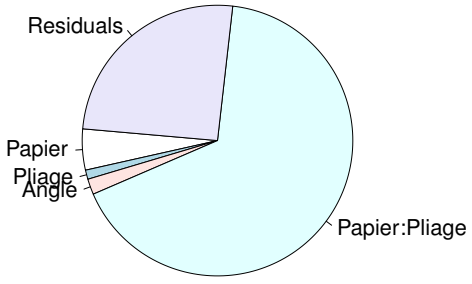
Les résultats de l'analyse de type I sont présentées dans la table qui suit ; le plan étant équilibré, l'ordre d'introduction des variables n'intervient pas. On garde ici le modèle P*T, l'angle n'apparaissant pas significatif. La qualité du papier n'a donc pas la même influence selon le type de pliage.

D		P	A	T
2160	1511	1	1	1
4596	3706	1	1	2
3854	1690	1	2	1
5088	4255	1	2	2
6520	4091	2	1	1
2130	3150	2	1	2
6348	4550	2	2	1
2730	2585	2	2	2

	Df	SS	F value	Pr(>F)
P	1	1718721	1.63	0.24
T	1	385641	0.367	0.56
A	1	654481	0.623	0.45
$P:T$	1	23386896	22.2	0.001
$P:A$	1	419904	0.4	0.54
$T:A$	1	73441	0.07	0.8
$P:T:A$	1	21025	0.02	0.89
Residuals	8	8392178		

TABLE III.3 – Données et analyse de type I sur le modèle D=P*A*T avec le logiciel R (`anova(lm(D~P*T*A))`). P=papier, T=type de pliage, A=angle de lancer, D=distance parcourue.

La contribution des facteurs peut s'illustrer par un camembert basé sur les SS, après une analyse éliminant les interactions trop faibles (ce genre de figure est à prendre avec précautions puisqu'elle dépend a priori de l'ordre dans lequel sont rentrés les facteurs ; ce n'est pas le cas ici où le plan est équilibré. Pour le choix des SS plutôt que $\sqrt{\text{SS}}$, voir par exemple l'argumentation de la fin du § II.2.4) :



III.4.5 Analyse de covariance

On est cette fois dans la situation où l'on a des régresseurs catégoriels et quantitatifs. Supposons que l'on en ait un de chaque ; le modèle de régression est

$$y_{ij} = \mu_i + a_i z_{ij} + u_{ij}.$$

Soit encore, sous forme vectorisée, $y = X\beta + u$ avec

$$y = \begin{pmatrix} y_{1.} \\ \vdots \\ y_{I.} \end{pmatrix}, \quad X = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} & z_{1.} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} & \mathbf{0} & z_{2.} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \dots & z_{I.} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu \\ a \end{pmatrix}.$$

où $\mathbf{0}$ et $\mathbf{1}$ sont des vecteurs de 0 et de 1, et $u_{i.}$ est le vecteur des u_{ij}. Le principe des tests est inchangé.

Exemple 1. Reprenons l'exemple de la page 56. On observe une variable explicative supplémentaire : Sodium. L'analyse du type I du modèle avec interaction puis l'analyse des coefficients du modèle additif donnent

	Df	Sum Sq	Pr(>F)
Viande	2	17692	7.1e-12
Sodium	1	18614	4.4e-13
Viande : Sodium	2	212	0.58
Residuals	48	9242	

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	75,74	8,7	1,6e-11
Divers	-1,66	4,5	0,717
Volaille	-49,8	4,7	2e-14
Sodium	0,2	0,02	2e-13

L'interprétation du 0,717 est que les viandes diverses n'ont pas d'apport calorique significativement différent du bœuf (le coefficient du bœuf est 0). L'interprétation du 0,58 est que l'effet calorique du sodium ne dépend pas de la viande. Si l'on regroupe ces deux classes on obtient après une analyse supplémentaire le modèle

$$\text{Calories}=75,2-49\ 1_{\text{Volaille}}+0,2\ \text{Sodium}+\text{ bruit},\qquad \widehat{\sigma}=13,6\qquad\qquad R^2=0,8.$$

Sur l'échantillon, la variable Sodium a une moyenne de 425 et un écart-type de 95.

Exemple 2. On s'intéresse à la relation entre l'activité sexuelle et la longévité chez les mouches [65, 47]. L'étude se base sur une expérience faite sur 5 groupes de 25 mouches mâles. Aux mouches du premier groupe, on a fourni une femelle vierge par jour, et à celle du deuxième groupe huit par jour ; les groupes 3 et 4 correspondent à la même expérience mais avec des femelles récemment inséminées (ce qui rend le rapport impossible) ; les mâles du groupe 5 sont seuls. Les variables sont

L : longévité en jours (réponse)

N : nombre de partenaires (0, 1 ou 8)

V : vierge (1), inséminée (0), aucun (-1, si nombre=0)

T : longueur du thorax du mâle en mm.

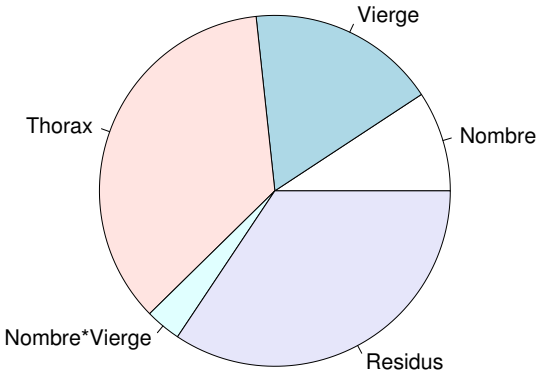
Sur les 9 combinaisons possibles pour N et V, seules 5 ont un sens. Pour éviter ce problème on peut retirer le 5e groupe de l'étude, ce qui fait un plan plus simple à 4 possibilités qui permet de tester un modèle additif. On va voir que le modèle additif sera refusé, ce qui fait on pourra passer à 5 groupes sans perturber le modèle.

Les résultats (logiciel R, tests de type I) de l'analyse de covariance pour le modèle final $L = N * V + T$ obtenu après élimination progressive des interactions non significatives sont (le 5e groupe a été éliminé) :

	Df	Sum Sq	F value	Pr(>F)
N	2	3542	16	6,8e-07
V	1	6675	60,4	3e-12
T	1	13633	123	$<$ 2,2e-16
$N:V$	1	1259	11,4	0,001
Residuals	119	13145		

	$V=1$	$V=0$
$N=8$	41	65
$N=1$	54	63,7
$N=0$	61	

Le tableau de droite exprime l'effet relatif en jours de la partie N∗V du modèle, par les prédictions obtenues à T fixe (valeur moyenne sur l'échantillon : 0,817 mm) dans les cinq groupes sur la base du modèle $L = N * V + T$ estimé sur l'échantillon complet, ce qui revient à $L = Groupe + T$: c'est un exemple typique de l'utilisation des coefficients pour l'interprétation. La contribution des facteurs peut s'illustrer par un camembert basé sur les Sum Sq :



III.4.6 Facteurs emboîtés (hiérarchisés, nested) en analyse de variance

Il se peut que la modalité prise par un des facteurs détermine celle prise par un autre, par exemple si les facteurs sont « ville d'origine » et « région d'origine ». Dans ce cas il n'est pas question d'introduire d'interaction mais on testera le modèle ne dépendant que de la région contre celui dépendant de la ville

$$y = \mu_r + u \quad \text{ou} \quad y = \mu_v + u.$$

Aspect pratique : Dans les données, les villes seront souvent numérotées à partir de 1 dans chaque région, c'est pour cela que les logiciels permettent de préciser que les facteurs « ville » et « région » sont hiérarchisés, ce qui permet de ne pas confondre des villes de même indice appartenant à des régions différentes. L'équation ci-dessus s'écrit alors $y_{rvk} = \mu_{rv} + u_{rvk}$ et un modèle paramétré sous contraintes est

$$y_{rvk} = \mu + \alpha_r + \gamma_{rv} + u_{rvk}, \quad \alpha_R = \gamma_{rV} = 0$$

un coefficient β_v n'ayant aucun sens. La commande `R` sera `lm(y~région+région/ville)`.

III.4.7 Modèles mixtes

Ces modèles ont été introduits dans le cadre de la régression au § II.3.5. Commençons par le modèle à effets aléatoires à un facteur :

$$y_{ik} = \mu + \alpha_i + u_{ik}, \quad u \sim \mathcal{N}(0, \sigma^2 I), \quad \alpha \sim \mathcal{N}(0, \sigma_\alpha^2 I)$$

et u et α sont indépendants. Les paramètres à estimer sont maintenant simplement μ, σ et σ_α. Ce modèle signifie que les y_{ik} forment un vecteur gaussien de moyenne μ et de covariance différente d'un multiple de l'identité, des corrélations apparaissant entre observations ayant même facteur i.

Par exemple, si l'on teste différents engrais (indice i) sur différentes cultures (indice j), l'utilisation du modèle mixte se justifie si l'on ne s'intéresse pas à la valeur explicite de l'interaction culture/engrais ; il pourra s'écrire

$$y_{ijk} = \mu_i + \nu_j + \alpha_{ij} + u_{ijk}, \quad u \sim \mathcal{N}(0, \sigma^2 I), \quad \alpha \sim \mathcal{N}(0, \sigma_\alpha^2 I). \tag{III.13}$$

Une faible valeur de σ_α indiquera que l'effet de l'engrais dépend peu des cultures.

Voir également l'exercice 5 p. 67 pour un autre exemple.

Bilan : Mixte contre fixe. Remarquons que le modèle mixte avec interaction aléatoires peut être identifié même si le modèle à effets fixes correspondant n'est pas identifiable (par manque d'observations, i.e. X déficiente) ; il est donc surtout intéressant pour tenir compte des interactions lorsqu'on n'a *pas assez de données* pour pouvoir les estimer explicitement ou bien que l'on ne cherche pas à les mesurer. Son

usage typique est de *tester la présence d'interactions* (ou plutôt de corrélations) dans un tel contexte : $H_1 =$« mixte » et $H_0 =$« fixe sans interaction ».

Dans le modèle longitudinal de la page 28, le modèle mixte a permis de constater que les enfants les plus grands à 12 ans sont ceux qui croissent le plus vite entre 12 et 14 ans ($r_{01} = 0,61$ significativement non nul).

Avec R, on peut faire des tests de modèles emboîtés par la commande `anova(.,.)` ou bien en utilisant `lrtest()` de la bibliothèque `lmtest` (il s'agit du test du rapport de vraisemblance, cf. § C.3.1). Il est préférable de forcer l'estimation par maximum de vraisemblance[4], l'estimateur par défaut étant généralement REML (REstricted Maximum Likelihood) ; ceci se fait en ajoutant l'option `REML=F` (cas `lmer`) ou `method="ML"` (cas `lme`).

Facteurs emboîtés. Prenons l'exemple de données longitudinales correspondant à la croissance de pommes sur des arbres. Les variables sont y le diamètre, t le temps, a l'arbre et p la pomme. Pomme et arbre sont emboîtés. Soit le modèle

$$\begin{aligned} y_{apk} =& \mu + \nu t_{apk} + (\alpha_a + \beta_a t_{apk}) + (\gamma_{ap} + \delta_{ap} t_{apk}) + u_{ijk}, \\ & (\alpha, \beta) \sim \mathcal{N}(0, R_A), \quad (\gamma, \delta) \sim \mathcal{N}(0, R_P), \quad u \sim \mathcal{N}(0, \sigma_\alpha^2 I). \end{aligned}$$

Pour a et p fixés, les t_{apk} sont donc les instants de mesure de diamètre de la pomme concernée. Comparer R_A et R_P revient à comparer les fluctuations d'un arbre à l'autre (en taille et en vitesse de croissance) aux fluctuations d'une pomme à l'autre à l'intérieur du même arbre. On pourra faire au choix

```
lme(y~t,random=~ t|a/p)
```
ou bien

```
lmer(y~t+(t|a)+(t|a:p))
```

p étant le numéro de la pomme dans l'arbre. L'utilisation de `lme` lorsqu'il y a plusieurs facteurs de groupe *non emboîtés* est semble-t-il impossible, p.ex. la commande `lme(y~1,random=list((~1|A),(~1|B)))` équivaut à `lme(y~1,random=list((~1|A/B)))` et correspond à $y_{ijk} = \mu + \alpha_i + \beta_{ij} + u_{ijk}$, c.-à-d. à l'emboîtement.

III.4.8 Réduction des interactions

Certains auteurs se défont des termes d'interaction en introduisant des variables censées les résumer :

$$y_{ijk} = \alpha_i + \beta_j + \sum_{q=1}^{Q} \gamma_q x_{ij}^q + u_{ijk}$$

où Q est petit et les x_{ij}^q sont des variables explicatives choisies à l'avance, censées représenter à elles seules les effets d'interaction. Par exemple i (resp. j) désigne la catégorie professionnelle (17 modalités) du père (resp. du fils), $x_{ij}^1 = S_i S_j$ où S_i est l'indice socioéconomique de la profession et $x_{ij}^2 = 1_{i=j} S_i^2$ (voir les détails au paragraphe § IV.2.2 où cette réduction est utilisée pour un modèle linéaire généralisé). On a ici $I + J + Q - 1$ paramètres au lieu de $I \times J$ pour le modèle avec interactions.

On aurait aussi pu considérer des termes d'analyse de covariance de la forme $\delta_j \gamma_i$ où la variable explicative γ_i peut être S_i, ou encore l'estimée de α_i dans le modèle additif, et δ_j le paramètre de pente à estimer.

III.4.9 Exercices

Exercice 1. Deux analyses de variance de sur les mêmes données vous fournissent les résultats suivants. Que pouvez-vous en dire ?

H_1	H_0	Pr
AB	$A+B$	0.4
$A+B$	A	0.4
A	cst	0.001

H_1	H_0	Pr
AB	$A+B$	0.4
$A+B$	B	0.4
B	cst	0.001

4. Cf. [28]. Voir § 1.3.1, § 1.5.1, § 2.2.4, § 4.1.2, ...

Exercice 2. Une analyse de variance donne les résultats suivants. Choisissez-vous le modèle A,B,$A+B$, le modèle complet, ou un autre ?

H_1	H_0	Pr
AB	$A+B$	0.001
$A+B$	B	0.4
B	cst	0.001

Exercice 3. On teste un engrais (facteur A). Malheureusement les champs engraissés sont principalement tous dans une région tandis que les champs témoins sont dans une autre. Il y a donc un facteur région B proche de A.

1. Quelle sera probablement la conclusion du test de significativité de A : $H_0 = B$ contre $H_1 = A+B$?
2. On oublie d'introduire le facteur de région. Comment teste-t-on le facteur A ? Quelle sera la conclusion si l'engrais a une influence significative ?
3. Quelle sera le résultat des tests précédents si la région a une influence significative et l'engrais est sans influence ?

Exercice 4. (Modèle mixte) On teste l'effet de deux médicaments $m = 1, 2$ sur différents sujets ; chaque sujet n'essaye qu'un médicament. La réponse est une variable mesurant l'amélioration de l'état de santé du sujet. Les sujets sont regroupés en G groupes (p.ex. selon l'âge) et l'on considère le modèle :

$$y_{mgk} = \alpha_m + \beta_{mg} + u_{mgk}, \quad \beta_{mg} \sim \mathcal{N}(0, \sigma_m^2), \quad u_{mgk} \sim \mathcal{N}(0, \sigma^2)$$

où y_{mgk} est la réponse du k-ième sujet du groupe g ayant pris le médicament m, et k va de 1 à K_{mg}. Il y a donc 3 paramètres de variance. Interpréter l'hypothèse « $\alpha_1 = \alpha_2, \sigma_1 < \sigma_2$ ».

Exercice 5. (Modèle mixte) Sur chacun des I sujets, on fait un prélèvement sanguin que l'on divise en n_i échantillons, envoyés à n_i des J laboratoires (plan incomplet). Chaque laboratoire divise l'échantillon en K et fait K mesures. On considère le modèle suivant pour les résultats

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + u_{ijk},$$

avec

$$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2),\ \beta_j \sim \mathcal{N}(0, \sigma_\beta^2),\ \gamma_{ij} \sim \mathcal{N}(0, \sigma_\gamma^2),\ u_{ijk} \sim \mathcal{N}(0, \sigma^2).$$

Combien ce modèle a-t-il de paramètres ? L'hypothèse « $\sigma_\gamma = 0$ » est acceptée. Interpréter les deux variances restantes.

Exercice 6. Soit le modèle linéaire avec expériences répétées :

$$y_{ij} = a + bx_i + u_{ij}, \quad i = 1, \dots I, \quad j = 1, \dots k_i.$$

On notera R l'erreur résiduelle obtenue après estimation de ce modèle aux moindres carrés.

Pour tester l'adéquation du modèle linéaire, on considère en parallèle le modèle d'analyse de la variance à un facteur à I modalités (comme si la variable x était catégorielle).

1. Écrire le modèle d'analyse de la variance à un facteur et expliciter la somme des carrés des erreurs résiduelles R_1.
2. On propose la statistique de test : $F = n_1(R - R_1)/(n_2 R_1)$.
   Donner n_1 et n_2. Quelle est l'hypothèse H_0 ? Quelle est la loi de F sous cette hypothèse ?
3. Expliciter le test à 5%.

Exercice 7. On considère le modèle à deux facteurs avec interaction sous les deux formes équivalentes (III.7) et (III.10) avec la convention (III.11)

1. Quelle est la particularité de la matrice γ_{ij} si $I = J = 2$?
2. Toujours si $I = J = 2$, donner 4 équations à quatre inconnues exprimant $(\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22})$ en fonction de $\mu, \alpha_1, \beta_1, \gamma_{11}$. Si $(\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}) = (0, 1, 3, 1)$, que valent $\mu, \alpha_i, \beta_j, \gamma_{ij}$?

Exercice 8. On considère l'analyse de la variance à un facteur avec le modèle sous les deux formes équivalentes (III.6) et (III.11).

1. Exprimer $\widehat{\mu}$ et $\widehat{\alpha}_i$ en fonction des $\widehat{\mu}_i$. Exprimer $\widehat{\mu}_i$ en fonction des observations.
2. Donner la matrice de covariance de $(\widehat{\mu}_1, \dots \widehat{\mu}_p)$; en déduire $\mathrm{Var}(\widehat{\mu})$ puis $\mathrm{Var}(\widehat{\alpha}_i)$. Simplifier ces expressions dans le cas où n_i ne dépend pas de i.

Exercice 9. On considère l'analyse de la variance à un facteur avec le modèle sous la forme

$$y_{ik} = \mu_i + u_{ik}, \qquad i = 1, \dots p, \qquad k = 1, \dots n_i.$$

1. Quelle est la variance de $\widehat{\mu}_i - \widehat{\mu}_j/2$? Donner un intervalle de confiance à de niveau α=5% pour $\mu_i - \mu_j/2$, centré en $\widehat{\mu}_i - \widehat{\mu}_j/2$, en fonction des n_k, de p et de $\widehat{\sigma}$. Vérifier que pour $(p, n_1, n_2, \widehat{\sigma}, \alpha) = (2, 35, 3, 0.15, 0.05)$ la largeur de l'intervalle pour $\mu_1 - \mu_2/2$ est 0,2 (cf. l'exercice 5 p. 52).
2. Montrer qu'on a l'intervalle de confiance de niveau α ($\chi^2_{n-p}(.)$ désigne le quantile du χ^2_{n-p}) :

$$(n-p)\widehat{\sigma}^2/\chi^2_{n-p}(1-\alpha/2) \leqslant \sigma^2_* \leqslant (n-p)\widehat{\sigma}^2/\chi^2_{n-p}(\alpha/2).$$

3. En déduire un intervalle de confiance de même niveau pour la variance de $\widehat{\mu}_i - \widehat{\mu}_j/2$. Vérifier que pour les valeurs considérées plus haut on obtient $\sigma_{\widehat{\mu}_1 - \widehat{\mu}_2/2} \in [0.054, 0.087]$ avec $\alpha = 5\%$.

Exercice 10. (Tests de type I) Réécrivons le modèle complet d'analyse de variance AB+C comme

$$\begin{aligned} AB + C &= c + A + B + C + A.B, \\ n_A n_B + n_C - 1 &= 1 + (n_A - 1) + (n_B - 1) + (n_C - 1) + (n_A - 1)(n_B - 1) \end{aligned}$$

écriture qui ne fait que déterminer cinq groupes dans les colonnes de X. On considère le nouveau modèle équivalent obtenu par orthogonalisation de chaque facteur aux précédents dans l'ordre donné par la syntaxe, le modèle $AB + C$ devenant :

$$AB + C = c + A + B^{\perp c,A} + A.B^{\perp c,A,B} + C^{\perp c,AB}.$$

On peut ensuite tester chaque facteur (A, B, A.B, C) en testant le modèle complet contre le modèle ci-dessus privé des colonnes correspondant au facteur. Montrer, en exploitant le lemme 10, que le numérateur de chaque statistique de Fisher est la norme de la projection de y sur l'espace correspondant, que ce test correspond au test de type I et que la somme des 4 numérateurs vaut $TSS - RSS$ (décomposition de la variance).

III.5 Un exemple de conclusion d'étude

Il s'agit des données 'CPS_85_Wages' disponibles sur `http://lib.stat.cmu.edu/datasets/`. On notera le travail d'analyse des régresseurs et des résidus. Voici la liste des variables et les conclusions de Therese Stukel (la réponse est le salaire horaire) :

WAGE (dollars per hour).
EDUCATION : Number of years of education.
SOUTH : 1=Person lives in South, 0=Person lives elsewhere.
SEX : 1=Female, 0=Male.
EXPERIENCE : Number of years of work experience.
UNION : 1=Union member, 0=Not union member.
AGE (years).
RACE : 1=Other, 2=Hispanic, 3=White.
OCCUPATION : 0=Other, 1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional.
SECTOR : 0=Other, 1=Manufacturing, 2=Construction.
MARIT : 0=Unmarried, 1=Married.

« The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience,

occupational status, region of residence and union membership. We wish to determine (i) whether wages are related to these characteristics and (ii) whether there is a gender gap in wages. Based on residual plots, wages were log-transformed to stabilize the variance. Age and work experience were almost perfectly correlated (r=.98). Multiple regression of log wages against sex, age, years of education, work experience, union membership, southern residence, and occupational status showed that these covariates were related to wages (pooled F test, $p < .0001$). The effect of age was not significant after controlling for experience. Standardized residual plots showed no patterns, except for one large outlier with lower wages than expected. This was a male, with 22 years of experience and 12 years of education, in a management position, who lived in the north and was not a union member. Removing this person from the analysis did not substantially change the results, so that the final model included the entire sample. Adjusting for all other variables in the model, females earned 81% (75%, 88%) the wages of males ($p < .0001$). Wages increased 41% (28%, 56%) for every 5 additional years of education ($p < .0001$). They increased by 11% (7%, 14%) for every additional 10 years of experience ($p < .0001$). Union members were paid 23% (12%, 36%) more than non-union members ($p < .0001$). Northerns were paid 11% (2%, 20%) more than southerns ($p = .016$). Management and professional positions were paid most, and service and clerical positions were paid least (pooled F-test, $p < .0001$). Overall variance explained was $R^2 = .35$.

In summary, many factors describe the variations in wages : occupational status, years of experience, years of education, sex, union membership and region of residence. However, despite adjustment for all factors that were available, there still appeared to be a gender gap in wages. There is no readily available explanation for this gender gap. »

IV

Régression linéaire généralisée

IV.1 Modèle linéaire généralisé

IV.1.1 Motivations. Définition

Dans bien des applications, les variables à expliquer ne varient pas dans tout $\mathbb{R}$ mais dans $\mathbb{R}_+$, $\mathbb{N}$ ou encore un intervalle d'entiers (cf. le « credit scoring » § I.2.3). Il est clair que le modèle gaussien est mal adapté à cette situation. Le modèle linéaire généralisé (GLM) spécifie que y_i est une variable aléatoire dont la loi est paramétrée par une combinaison linéaire des régresseurs $x_i\beta$, par exemple $y_i \sim \mathcal{P}(x_i\beta)$.

En pratique la situation est la suivante : on dispose de données y et X (réponses et variables explicatives) ; il faut alors spécifier une famille $(P_\theta)_{\theta\in\mathbb{R}}$ de distributions de probabilité à un paramètre réel θ ainsi qu'une fonction réelle $\eta \mapsto r(\eta)$, dont l'inverse est appelé fonction de lien, qui fait le lien entre le paramètre θ et $x\beta$. Le modèle est alors

$$y_i \sim P_{\theta_i}, \quad i = 1, \dots n \tag{IV.1}$$

$$E_{\theta_i}[Y] = r(x_i\beta). \tag{IV.2}$$

Cette dernière équation présuppose que l'application $\theta \mapsto E_\theta[Y]$, est bijective ce qui sera toujours le cas, car on manipulera essentiellement des *familles exponentielles* dont θ est le paramètre naturel (le facteur de y dans l'exponentielle). On voit que modèle linéaire gaussien rentre dans ce cadre avec la famille $\mathcal{N}(\theta, \sigma^2)$ et $r(\eta) = \eta$.

Choix de la famille P_θ. Les logiciels proposent typiquement les familles

1. Valeurs réelles positives
   - Gamma
   - Inverse gaussienne
2. Valeurs entières positives non bornées
   - Poisson
3. Valeurs entières positives dans un intervalle
   - Binomiale

La distribution binomiale négative est également proposée mais semble peu utilisée (cf. [15] § 11.2).

Choix de r. Le choix par défaut proposé par les logiciels est $r(\eta) = E_\eta[Y]$, ce qui conduit à $\theta_i = x_i\beta$; ce choix permet une estimation numériquement robuste, et conduit à des valeurs réalistes indépendamment de x_i (i.e., p.ex. comprises entre 0 et 1 si la loi est binomiale). Pour comprendre les implications du choix de r, précisons la paramétrisation des modèles. Ces familles ont toutes une densité de la forme

$$f(y;\theta,\varphi) = \exp\left\{\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right\} \tag{IV.3}$$

où θ est le paramètre de localisation, et φ un paramètre supplémentaire de dispersion, appelé paramètre de nuisance, qui devra être estimé mais ne dépend pas des variables explicatives. C'est le jeu d'équations (IV.1, IV.2, IV.3) qui définit le cadre standard des modèles linéaires généralisés. Si φ est fixé, c'est une famille exponentielle, c'est pour cela qu'on appelle (IV.3) famille exponentielle à un paramètre de nuisance. On vérifie par le calcul que le lien canonique est $r_c = b'$.

Il est facile de voir qu'il existe une fonction V, dite fonction variance, telle que

$$Var(y) = \varphi V(\mu), \quad \mu = E[y] = b'(\theta)$$

(en fait $V = b'' \circ (b')^{(-1)}$) si bien qu'on aura l'équation

$$Var(y_i) = \varphi V(r(x_i\beta))$$

qui est une façon supplémentaire de relier r aux données.

Nous résumons et précisons cette discussion dans le tableau suivant, en notant r_c le lien canonique :

Loi	Supp.	densité	μ	σ^2	θ	φ	$V(\mu)$	$r_c(\eta)$
$\mathcal{B}(m,p)$	$[0..m]$	$C_m^y p^y (1-p)^{m-y}$	mp	$mp(1-p)$	$\log\left(\frac{p}{1-p}\right)$	1	$\mu(1-\frac{\mu}{m})$	$\frac{1}{1+e^{-\eta}}$
$\mathcal{P}(\mu)$	$\mathbb{N}$	$\mu^y e^{-\mu}\frac{1}{y!}$	μ	μ	$\log\mu$	1	μ	e^η
$\mathcal{N}(\mu,\sigma^2)$	$\mathbb{R}$	$\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	μ	σ^2	1	η
$\Gamma(p,\beta)$	$\mathbb{R}_+$	$y^{p-1}\beta^{-p}e^{-y/\beta}/\Gamma(p)$	βp	$\beta^2 p$	$-1/\mu$	p^{-1}	μ^2	$-1/\eta$
$\mathrm{IG}(\mu,\lambda)$	$\mathbb{R}_+$	$\exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\}\sqrt{\frac{\lambda}{2\pi y^3}}$	μ	μ^3/λ	$-1/\mu^2$	$2/\lambda$	$\mu^3/2$	$1/\sqrt{-\eta}$

Les fonctions de lien typiquement proposées par les logiciels sont (Φ désigne la fonction de répartition de la Gaussienne) :

Lien	$\mu = r(\eta)$
identité	η
logarithme	e^η
logit	$1/(1+e^{-\eta})$
loglog complémentaire	$1-\exp(-e^\eta)$
probit	$\Phi(\eta)$
puissance	$\eta^{1/\alpha}$

Modèles mixtes. La théorie des modèles mixtes, qui consiste à considérer certain paramètres, comme aléatoires existe de la même façon pour les GLM. Nous n'en traiterons pas spécifiquement. Un exemple est donné à l'exercice 4 p. 77.

IV.1.2 Exercices

On pourra préférer lire la partie suivante avant de faire ces exercices.

Exercice 1. Montrer que les modèles suivants sont des modèles linéaires généralisés :

1. $y_i = \begin{cases} 1 \text{ si } x_i + az_i + b^3\log(x_i) + e_i \geqslant 0 \\ 0 \text{ sinon} \end{cases}$
2. $y_i \sim \mathcal{N}(\alpha_0 x_i^{\beta_1} z_i^{\beta_2}, \sigma^2)$
3. $y_i = \begin{cases} \mathcal{B}(1,p) \text{ si } x_i = 0 \\ \mathcal{B}(1,q) \text{ si } x_i = 1. \end{cases}$

La paire (x_i, z_i) est le régresseur pour l'individu i et les e_i sont i.i.d de fonction de répartition $1/(1+e^{-t})$.

On explicitera φ, β , les fonctions r et b ainsi que les régresseurs à considérer.

Exercice 2. On considère le modèle poissonnien $y \sim \mathcal{P}(e^{x\beta})$. Écrire l'équation satisfaite pour l'estimateur au maximum de vraisemblance pour β.

Exercice 3. La loi binomiale négative $\mathcal{B}_-(\mu, \alpha)$ sur $\mathbb{N}$ donne à l'entier n la probabilité

$$p_{\mu,\alpha}(n) = \frac{\Gamma(\alpha+n)}{n!\Gamma(\alpha)} \frac{\mu^n \alpha^\alpha}{(\mu+\alpha)^{n+\alpha}}.$$

Sa moyenne est μ et sa variance $\mu + \mu^2/\alpha$. Pour α entier, son interprétation est la suivante : soit T l'instant du α-ième succès dans un Bernouilli de probabilité $p = \alpha/(\mu+\alpha)$; alors $T - \alpha$ suit une loi $\mathcal{B}_-(\mu, \alpha)$.

1. Montrer que pour tout n, $p_{\mu,\alpha}(n)$ tend vers une limite (que l'on identifiera) quand $\alpha \to \infty$.
2. α est fixé. Donner θ, $b(\theta)$, et $b'(\theta)$. Plusieurs choix sont possibles pour θ ; on fera celui qui conduit au paramètre de la loi de Poisson quand $\alpha \to \infty$.

IV.2 Exemples

IV.2.1 Variable de Bernoulli : le modèle logistique

Reprenons l'exemple du test de l'insecticide :

$$y = \begin{cases} 1 \text{ si la blatte meurt} \\ 0 \text{ sinon} \end{cases}$$

$$x = (\text{dose, produit, souche}) = (z, j, s)$$

avec

- dose : variable quantitative
- produit : variable catégorielle à 3 modalités
- souche : variable catégorielle à 4 modalités.

Le modèle linéaire généralisé sans interaction naturel est

$$\boxed{y \sim \mathcal{B}(1, r(x\beta))} \qquad \beta \in \mathbb{R}^7.$$

Le lien sera en pratique choisi parmi « logit » (modèle logistique, lien canonique) ou « probit » ce qui donne

$$P(y=1) = \frac{1}{1+e^{-x\beta}} \quad \text{ou} \quad P(y=1) = \Phi(x\beta).$$

Certains logiciels proposent également de mettre un seuil :

$$r(z) = c + (1-c) r_0(z)$$

où $r_0(.)$ est « logit » ou « probit ». Ce seuil permet d'autoriser la contrainte $P(y=1) \geqslant c$ quel que soit x.

Expériences de Bernoulli répétées. Reprenons l'exemple précédent mais supposons qu'on ait fait des lots de blattes où toutes les blattes du même lot ont les mêmes conditions d'expérience (même x_i). Il est naturel de rassembler les résultats lot par lot, sans distinguer les blattes.

Pour chaque lot, désignons par m le nombre de blattes, X la valeur commune du régresseur et Y le nombre de mort. On peut considérer que l'on observe les variables $(Y_l, X_l, m_l)_{1 \leqslant l \leqslant L}$ où L est le nombre de lots ($\sum_l m_l = n$) et la distribution de Y_l est :

$$\boxed{Y_l \sim \mathcal{B}(m_l, r(X_l\beta))}$$

Exemple. On fait tester 7 marques de corn flakes par 100 personnes. Les tests sont faits par paires : chacun fait 21 expériences consistant à goûter deux marques différentes et dire laquelle il trouve plus croustillante [37]. Dans le tableau suivant la case (i, j) indique combien de testeurs ont trouvé i plus croustillante que j :

	1	2	3	4	5	6	7
1	0	39	64	40	61	76	46
2	61	0	65	59	55	85	60
3	36	35	0	31	25	41	35
4	60	41	69	0	41	80	28
5	39	45	75	59	0	71	37
6	24	15	59	20	29	0	18
7	54	40	65	72	63	82	0

On peut proposer le modèle $y_{ij} \sim \mathcal{B}(100, r(\beta_i - \beta_j))$ où β_i est la « croustillance » de la i-ième variété, et r doit satisfaire $r(0) = 0,5$ ce qui est bien le cas du lien canonique. Comme seules les différences interviennent, on peut poser $\beta_7 = 0$ et il n'y a que 6 paramètres, $x_{ij} \in \{-1, 0, 1\}^6$. Les intervalles de confiance obtenus pour les β_i conduisent à un regroupement en 3 classes où les β_i ne sont pas significativement distincts : $\{\beta_2, \beta_7\}$, $\{\beta_1, \beta_4, \beta_5\}$, $\{\beta_3, \beta_6\}$ (par ordre de croustillance décroissante ; on peut faire le test sous R avec la commande `lht` de la bibliothèque `car`). Notons que l'on retrouve dans cet exemple un cas de non-monotonicité des tests au sens où la p-value associée à H_0 :« $\beta_3 = \beta_6$ » est inférieure à celle de H_0 :« $\beta_2 = \beta_7, \beta_1 = \beta_4 = \beta_5, \beta_3 = \beta_6$ ».

► Il arrive souvent que les données soient fortement déséquilibrées au sens où $y_i = 0$ massivement dans l'échantillon. Les statisticiens ont observé qu'il est alors fructueux d'utiliser des méthodes de **rééquilibrage** de l'échantillon [5].

► Mentionnons également la possibilité de faire des **modèles mixtes**, voir l'exercice 4 p. 77. Sous R, ils sont traités par la fonction `glmer` de la bibliothèque `lme4`

► Il a été remarqué depuis longtemps que la régression logistique et l'analyse discriminante poursuivent essentiellement le même but. Il est généralement admis que l'hypothèse de distribution gaussienne pour les variables explicatives est importante pour que l'analyse discriminante donne de bon résultats ; en particulier, en présence de variables catégorielles la régression logistique devrait être meilleure [68].

► Une autre interprétation du modèle logistique par le modèle inversé : Supposons que y_i soit tiré selon une loi de Bernoulli $\mathcal{B}(1, p)$ et que conditionnellement à $y_i = \varepsilon$, $x_i \sim \mathcal{N}(\mu_\varepsilon, R)$. Les paramètres du modèle sont donc (p, μ_0, μ_1, R). On montre alors facilement avec la formule de Bayes que $\log P(y_i = 0|x_i) = \alpha + x_i\beta$ pour un certain scalaire α et un certain vecteur β qui s'expriment simplement en fonction des paramètres.

IV.2.2 Modèle poissonnien

Premier exemple : On compte sur plusieurs années, en chaque saison, le nombre d'accidents sur certaines routes

y_i = nombre d'accidents
x_i = (nombre de voies sur la route, saison, investissement annuel en entretien de la route)

Le premier régresseur a 2 modalités et le deuxième régresseur en a 4. Le modèle naturel est poissonnien, ce qui donne avec lien canonique :

$$\boxed{y \sim \mathcal{P}(\mu), \quad \log(\mu) = x\beta}$$

(avec ici $\beta \in \mathbb{R}^6$), soit encore $E[y] = e^{x\beta}$. C'est un modèle log-linéaire. Il se peut que le lien identité soit mieux adapté que le lien log-linéaire qui implique un effet multiplicatif des facteurs.

Si y_i est le nombre de cas de grippe dans le département i, la réponse naturelle serait y_i/p_i, où p_i est la population du département, ce qui ne peut se faire si l'on veut conserver le modèle poissonnien. La façon usuelle de prendre p en compte est d'introduire $\log(p)$ en « offset » (prédicteur pour lequel β est connu) : $E[y] = e^{\log(p)+x\beta}$.

La contrainte de variance égale à la moyenne pour la loi de Poisson peut être levée en utilisant une loi **binomiale négative** qui, elle, possède un paramètre de localisation *et* un paramètre de nuisance qui modifie la dispersion (fonction `glm.nb` de R). Toutefois la dépendance en φ est plus compliquée que celle de la formule (IV.3), en particulier le lien canonique dépend du paramètre de nuisance (cf. [15] § 11.2). Noter que la variance est, pour une telle variable, toujours supérieure la moyenne.

Tables de contingence. Le modèle poissonnien avec lien logarithmique est également utilisé pour l'analyse des tables de contingence (n_{ijk}), où n_{ijk} est la réponse, les régresseurs sont catégoriels, et les modèles sont fabriqués avec certaines interactions, par exemple

$$n_{ijk} \sim \mathcal{P}(e^{\mu+\alpha_i+\beta_j+\gamma_k+\delta_{ij}}) \tag{IV.4}$$

avec indépendance des n_{ijk} (par exemple le nombre d'accidents à un carrefour, $i =$« jour/nuit », $j =$« conducteur jeune/âgé », etc.). La probabilité pour un individu de tomber dans la case i, j, k est la proportion d'individus qu'on trouvera dans cette case si l'on répète l'expérience un grand nombre de fois, soit $E[n_{ijk}]/E[n]$, dont le logarithme vaut donc

$$\boxed{\log p_{ijk} = c + \alpha_i + \beta_j + \gamma_k + \delta_{ij}} \qquad c = -\log \textstyle\sum_{p,q,r} e^{\alpha_p+\beta_q+\gamma_r+\delta_{pq}}$$

ce qui est une réécriture du modèle. On a donc l'interprétation :

$$\{p_{ijk}\} \textit{ représente la distribution des variables } i, j, k \textit{ dans l'échantillon}$$

(« dans l'échantillon » signifie ici « parmi les fauteurs d'accident ») de même que l'estimateur habituel $\widehat{p}_{ijk} = n_{ijk}/n$, qui correspond ici au modèle avec toutes les interactions.

Le lien logarithmique permet donc de représenter l'indépendance comme la nullité de certains coefficients d'interaction, ici $\delta_{ij} = 0$. En effet dans ce cas la probabilité d'observer (i, j, k) est proportionnelle à $e^{\alpha_i} e^{\beta_j} e^{\gamma_k}$ exprimant ainsi l'indépendance des trois facteurs. Dans le cas contraire on conclurait par exemple à une surreprésentation des jeunes dans les accidents nocturnes. Le modèle (IV.4) présuppose l'indépendance de (i, j) et k. Les coefficients $\alpha_i, \beta_j, \gamma_k$ ne font que refléter la proportion de chaque modalité dans l'échantillon ; l'information réelle se trouve dans les interactions.

Contrairement à ce qui a été vu jusqu'à présent, la régression a ici pour objet de conclure à des liens entre les variables explicatives ; ces dernières se présentent donc naturellement comme des variables aléatoires, tout du moins pour ce qui est de l'interprétation des résultats.

La cohérence du modèle poissonnien alors que le vecteur des n_{ijk} suit une loi multinomiale, provient du résultat suivant :

> *Soient $n_1, \dots n_K$ des v.a. indépendantes de loi de $\mathcal{P}(\mu_k)$, alors, conditionnellement à $\sum n_k = n$, la loi de $n_1, \dots n_K$ est multinomiale $\mathcal{M}(n; p_1, \dots p_K)$, $p_k = \mu_k/\lambda$, $\lambda = \sum_k \mu_k$.*

C'est immédiatement vérifié en utilisant que $n \sim \mathcal{P}(\lambda)$:

$$P(n_1, \dots n_K | n) = \frac{\mu_1^{n_1} e^{-\mu_1}}{n_1!} \dots \frac{\mu_K^{n_K} e^{-\mu_K}}{n_K!} \Big/ \frac{\lambda^n e^{-\lambda}}{n!} = \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \dots p_K^{n_K}$$

Notons que réciproquement : *Si conditionnellement à leur somme n les v.a. $n_1, \dots n_K$ suivent une loi multinomiale $\mathcal{M}(n; p_1, \dots p_K)$ et si n suit une loi $\mathcal{P}(\lambda)$, alors, les v.a. n_k sont indépendantes de loi $\mathcal{P}(\lambda p_k)$.* Le calcul est le même.

Exemple. Pour étudier la mobilité sociale, M. Hout [51] considère une table de contingence (p_{ij}) où i (resp. j) désigne la catégorie professionnelle (17 modalités) du père (resp. du fils). Il exploite l'idée

présentée au § III.4.8 : $Q = 5$ avec $x^1_{ij} = S_i S_j$ où S_i est l'indice socioéconomique de la profession (lié au prestige, au salaire, etc.), $x^2_{ij} = 1_{i=j} S^2_i$, et les trois autres sont fabriqués dans le même esprit avec d'autres indices :

$$\log p_{ij} = c + \alpha_i + \beta_j + \sum_{q=1}^{5} \gamma_q x^q_{ij}.$$

Avec R :

```
glm(nb~CatPere+CatFils+X1+X2+X3+X4+X5,fam=poisson)
```

En présence d'interaction significative, on conclura à la présence de lien significatif entre la profession du père et celle du fils.

EXEMPLE. [25] On fait faire le test suivant à 124 étudiants :

Le point z se trouve dans le quatrième quadrant (i.e. $-\pi/2 < \arg z < 0$) du plan complexe. Démontrer ou réfuter les propositions suivantes

(A) $arg(z + \bar{z}) = 0$ *(B)* $arg(z - \bar{z}) = \pi/2$
(C) z^2 *est dans le troisième quadrant* *(D)* $-\pi < arg(z/\bar{z}) < 0$

On s'intéresse aux variables « Résultat », « Sexe » et « Question » et l'on résume les données dans le tableau suivant :

	Hommes				Femmes			
	A	B	C	D	A	B	C	D
Correct	77	58	65	47	21	17	19	10
Incorrect	10	28	17	24	5	9	5	9
Non-réponse	7	8	12	23	4	4	6	11
Total	94	94	94	94	30	30	30	30

	Df	P(>|Chi|)
Résultat	2	3e-42
Question	3	1.00
Sexe	1	5e-32
Rés : Ques	6	4.5e-07

Le tableau de droite est la table d'analyse de variance ascendante donnée par R. Après élimination des facteurs non-significatifs, on a gardé le modèle $R*Q+S$. Il n'y a pas de différence significative de résultat entre les sexes, mais il existe une corrélation entre la question et le résultat : les questions n'ont pas même difficulté. Ceci se confirme bien sur les données.

Comme il s'agit d'une analyse ascendante, les chiffres $3e$-42, 1.00 et $5e$-32 correspondent à des test contre un modèle additif croissant (indépendance des facteur). Le 1.00 signifie simplement qu'il y a autant de questions de chaque type dans l'échantillon tandis que $3e$-42, et $5e$-32 signifient qu'il n'y a pas le même nombre d'hommes que de femmes et que les résultats ne sont pas uniformément distribués (il y a beaucoup plus de réponses correctes). La nature du plan fait qu'il ne peut pas y avoir d'interaction significative Sexe : Question (la variable Sexe ne donne aucune information sur la question posée).

IV.2.3 Modèle à variable catégorielle ordonnée ; la variable latente

Exemple :

$$y = \begin{cases} 3 \text{ si l'individu pratique du sport tous les jours} \\ 2 \text{ si l'individu pratique du sport au moins toutes les semaines} \\ 1 \text{ sinon} \end{cases}$$

$x =$ (âge, type d'études suivies, ...)

Introduisons la variable z_i qui représente l'envie qu'a l'individu de faire du sport et postulons le modèle :

$$z = x\beta + u$$
$$y = k \quad \text{si} \quad a_{k-1} < z \leqslant a_k, \quad a_0 = -\infty, \quad a_K = +\infty.$$

Soit

$$\boxed{P(y \leqslant k) = F(-x\beta + a_k)}$$

où $F(.)$ est la fonction de répartition de u. Le lien logit revient à prendre $F(x) = (1 + e^{-x})^{-1}$. Si $K = 2$ et u est gaussienne, on retrouve le modèle logistique avec lien probit (quitte à changer r en $1 - r$, ce qui revient à changer y en $1 - y$). Il faudra estimer β mais aussi les a_k ; ces derniers jouent le rôle d'intercept (coefficient de la constante), et x_{i1} est la première variable explicative effective (et non 1). En pratique F sera la fonction inverse du lien logit ou probit.

Noter que ce modèle ne rentre pas rigoureusement dans le formalisme des modèles linéaires généralisés donné par (IV.1, IV.2, IV.3).

L'analyse peut être réalisée sous R avec la fonction `polr()` de la bibliothèque `MASS` et les tests avec la fonction `Anova()` de la bibliothèque `car`.

IV.2.4 Modèle à variable catégorielle non-ordonnée (multinomial logit).

Ce modèle est appelé également « softmax regression » dans le milieu de l'apprentissage. Soit un sondage donnant les variables suivantes :

$$y_i = \text{Distraction préférée du samedi soir} \begin{cases} 1 = \text{spectacle} \\ 2 = \text{télévision} \\ 3 = \text{visite d'amis} \\ 4 = \text{autres} \end{cases}$$

$$x_i = (\text{âge, sexe, } \ldots)$$

On pourra utiliser le modèle avec ici $K = 4$:

$$\boxed{P(y = j) = \frac{\exp(x\beta_j)}{\sum_{k=1}^{K} \exp(x\beta_k)}, \qquad \beta_K = 0}$$

On peut toujours se ramener à $\beta_K = 0$ quitte à remplacer les β_j par $\beta_j - \beta_K$, ce qui ne change rien par ailleurs. La condition $\beta_K = 0$ évite donc la surparamétrisation. Ce modèle ne rentre pas tout-à-fait dans le cadre théorique mais généralise le modèle binomial logistique. Il est traité par la fonction `vglm` du package `vgam` de R, et par la fonction `catmod` de SAS. Il y a ici aussi une interprétation en termes de variable latente : Tirer des v.a. $z_j \sim \mathcal{E}(x\beta_j)$ et choisir le j correspondant à la plus petite (exercice 6 p. 78).

IV.2.5 Exercices

Exercice 1. On reprend l'exemple du § IV.2.1 (blattes). Quelle est la dimension de β si l'on suppose une interaction entre la dose et la souche ? entre le produit et la souche ?

Exercice 2. Vérifier la propriété de modèle inversé présentée à la dernière remarque du § IV.2.1 p. 74.

Exercice 3. On veut savoir si la présence d'un agent à un certain carrefour améliore la circulation. Pour cela on compte plusieurs fois le nombre de voitures qui attendent au carrefour en présence et en absence d'agent. On recueille alors un tableau de données (n_i, a_i, s_i) où n_i est le nombre de voitures et a_i vaut 0 s'il n'y a pas d'agent et 1 sinon. A été ajoutée la variable s_i qui est le sexe de l'agent pour voir si cette variable a de l'influence sur l'efficacité.

Proposer un modèle linéaire généralisé pour ces données. Quelle est la dimension de β ? Comment tiendriez-vous compte de l'heure si on l'avait mise dans les données ?

Exercice 4. (**Modèle logistique mixte** [53]) On a suivi les accouchements d'un certain nombre de femmes afin de mesurer l'importance d'un risque génétique sur les fausses couches. On a le tableau suivant : On dispose en réalité des variables suivantes, où i est l'indice de la femme et j le numéro

	< 35 ans		$\geqslant 35$ ans	
	vivant	fausse couche	vivant	fausse couche
sans risque	144	18	7	1
avec risque	121	57	8	5

d'accouchement pour cette femme :

$$y_{ij} = \begin{cases} 0 & \text{fausse couche} \\ 1 & \text{sinon} \end{cases} \qquad z_{ij} = \begin{cases} 0 & \text{âge} < 35 \text{ ans} \\ 1 & \text{sinon} \end{cases} \qquad h_i = \begin{cases} 0 & \text{risque absent} \\ 1 & \text{sinon.} \end{cases}$$

1. Proposer à partir du tableau un test classique pour voir, pour chaque classe d'âges, si les chances de fausse couche en absence ou en présence de risque sont les mêmes.
2. Soit le modèle pour la probabilité p_{ij} d'une fausse couche au j-ième accouchement de la i-ième femme :

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \lambda + az_{ij} + bh_i.$$

   De quel type de modèle s'agit-il ? Écrire la vraisemblance des réponses y_{ij}.
3. On propose le modèle mixte suivant :

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \lambda_i + az_{ij} + bh_i.$$

   où les λ_i sont des v.a. normales i.i.d $\mathcal{N}(0, \sigma^2)$. Pourquoi n'est-il pas raisonnable de considérer le modèle (non-mixte) où les λ_i sont des paramètres ? Interpréter ce modèle, en particulier concernant la présence éventuelle d'autres facteurs inconnus. Interpréter les tests « $a = 0$ », « $b = 0$ », et « $\sigma = 0$ ».

Exercice 5. On reprend l'exercice 3 p. 30. La difficulté est en fait mesurée par une note entre 1 et 3. Proposer un modèle linéaire généralisé mixte (s'inspirer de l'exercice précédent).

Exercice 6. Vérifier que le modèle à variable catégorielle non-ordonnée correspond au modèle à variables latentes annoncé.

IV.3 Estimation de β^* et φ_*

Nous présentons ici les idées essentielles qui guident l'estimation des paramètres. Pour plus de détails nous renvoyons à voir [15] ou [9].

IV.3.1 L'estimateur du maximum de vraisemblance

Dans tout ce paragraphe on suppose φ connu. En vertu de (IV.3), pour estimer β^* au maximum de vraisemblance, il faut maximiser en β

$$\mathcal{L}(\beta) = \varphi^{-1} \sum_{i=1}^{n} y_i\theta_i - b(\theta_i), \qquad b'(\theta_i) = r(x_i\beta). \tag{IV.5}$$

On a abusivement oublié les termes $c(y_i, \varphi)$ qui ne jouent aucun rôle. Précisons tout de suite que pour des fonctions r arbitraires, il peut très bien y avoir des maximums locaux. La dérivée est

$$\mathcal{L}'(\beta) = \varphi^{-1} \sum_{i=1}^{n} (y_i - b'(\theta_i)) \frac{\partial \theta_i}{\partial \beta}$$

et la relation liant θ_i à β donne $b''(\theta_i)\frac{\partial\theta_i}{\partial\beta} = r'(x_i\beta)x_i$. En substituant, et en introduisant la fonction variance $V(\mu) = b''(\theta)$, on trouve

$$\mathcal{L}'(\beta) = \varphi^{-1}\sum_{i=1}^{n}\frac{y_i-\mu_i}{V(\mu_i)}r'(x_i\beta)x_i, \quad \mu_i = r(x_i\beta).$$

On voit que l'annulation de cette fonction de β est un problème a priori assez compliqué. Si l'on introduit les variables

$$\check{x}_i = r'(x_i\beta)\ x_i$$
$$D = \mathrm{diag}(V(x_1),\dots V(x_n))^{-1}$$

on a

$$\varphi\mathcal{L}'(\beta) = \check{X}^T D(y-\mu). \tag{IV.6}$$

Divers algorithmes bien établis existent pour annuler cette fonction compliquée de β ; on verra le plus utilisé au § IV.3.3.

Cas du lien canonique. Si $r = b'$, l'équation devient $X^T(y-\mu) = 0$. La résolution de (IV.5) ne pose aucun problème fondamental car la fonction à maximiser est concave en β.

IV.3.2 Propriétés asymptotiques

On s'intéresse à l'asymptotique quand le nombre n d'observations (x_i, y_i) tend vers l'infini.

L'estimateur $\widehat{\beta}_n$ de β^* est l'estimateur au maximum de vraisemblance. Il est impossible de montrer que les hypothèses nécessaires à l'application des théorèmes concernant les propriétés asymptotiques du maximum de vraisemblance sont vérifiées en toute généralité. Supposons-les satisfaites et appliquons les résultats de l'annexe C. On a alors la convergence presque sûre de $\widehat{\beta}_n$ vers β^*. Si l'on pose (matrice d'information de Fisher)

$$\mathcal{I}_n = E\left[\mathcal{L}'_n(\beta^*)\mathcal{L}'_n(\beta^*)^T\right] = \varphi_*^{-1}\check{X}^T D\check{X} \tag{IV.7}$$

(cf. formule (IV.6)) on a alors normalité asymptotique

$$\mathcal{I}_n^{1/2}(\widehat{\beta}_n - \beta^*) \longrightarrow \mathcal{N}(0, Id).$$

Dans ces résultats, on peut remplacer $\mathcal{I}_n$ par $\widehat{\mathcal{I}}_n$, matrice calculée comme dans la formule (IV.7) sauf que les normalisations sont faites avec les paramètres estimés.

IV.3.3 Estimation de φ_* et β^*

La consistance de $\widehat{\beta}$ implique (sous certaines hypothèses) que

$$\widehat{\varphi} = \frac{1}{n}\sum_i V(\widehat{\mu}_i)^{-1}(y_i - \widehat{\mu}_i)^2 \tag{IV.8}$$

est un estimateur consistant de φ_*. Pour avoir une formule analogue au cas linéaire, on pourra préférer remplacer n par $n-p$ au dénominateur. On vérifie cependant facilement sur des simulations que cet estimateur est fréquemment assez mauvais ; il est meilleur de faire une estimation au maximum de vraisemblance (ce qui est facile car il s'agit de maximiser une fonction d'une seule variable).

Un algorithme d'estimation de β^*. L'algorithme de Newton pour la maximisation de $\mathcal{L}(\beta)$ est : $\beta_{\text{new}} = \beta - \mathcal{L}''(\beta)^{-1}\mathcal{L}'(\beta)$.

Malheureusement la matrice de dérivée seconde est généralement difficile à calculer. On préfère la remplacer par l'approximation $-\widehat{\mathcal{I}}_n$ (cf. § C.1), d'où l'algorithme

$$\beta_{\text{new}} = \beta + \left(\check{X}^T\check{D}X\right)^{-1}\check{X}^T D(\check{y} - \check{\mu})$$

où tout est calculé avec la valeur courante de β.

IV.4 Tests et analyse de déviance

IV.4.1 Déviance.

La déviance est utilisée comme mesure d'adéquation du modèle aux données. Elle vaut

$$D(\widehat{\beta}) = 2\varphi(\mathcal{L}_s - \mathcal{L}(\widehat{\beta}))$$

où $\mathcal{L}_s$ est la vraisemblance du modèle saturé, c-à-d du modèle avec un paramètre différent pour chaque donnée. Pour ce modèle, $\mu_i = y_i$ et donc :

$$\mathcal{L}_s = \varphi^{-1} \sum_{i=1}^{n} y_i\theta_i - b(\theta_i), \qquad b'(\theta_i) = y_i.$$

Noter que $D(\widehat{\beta})$ ne dépend pas de φ ; dans le cas du modèle normal, elle n'est autre que le RSS. Cette quantité difficile à interpréter n'a d'intérêt que purement indicatif. La déviance normalisée, $\varphi^{-1}D(\widehat{\beta})$, est plus étroitement liée à la vraisemblance et donc intervient naturellement dans les tests.

IV.4.2 Tests

On utilise les méthodes générales proposées à l'appendice C en exploitant les expressions obtenues pour la vraisemblance et la matrice d'information de Fisher (IV.7). En particulier, comme la différence de déviance normalisée entre deux modèles de même φ n'est autre que le logarithme du rapport de vraisemblance, on a asymptotiquement sous H_0 : $(D_0 - D_1)/\varphi \sim \chi^2_{p_1-p_0}$ (cf. § C.3.1), d'où le test

$$\frac{D_0 - D_1}{\varphi} \gtrless \chi^2_{p_1-p_0}(1-\alpha).$$

Pour les modèles pour lesquels φ n'est pas connu, il sera en pratique estimé sur le modèle le plus compliqué (supposé valide), et par analogie avec le cas linéaire, on fait le test :

$$\frac{D_0 - D_1}{(p_1 - p_0)\widehat{\varphi}} \gtrless f_{p_1-p_0,n-p_1}(1-\alpha).$$

Ces tests étant basés sur les résultats asymptotiques, il est plus prudent, si n est petit, d'estimer directement (par simulation d'échantillons sous H_0) les quantiles désirés de la loi sous H_0 de la statistique considérée (cf. § C.3.4).

Noter que lorsque φ est connu (modèle binomial ou poissonnien) le test du χ^2 reste valide même si H_1 donne une déviance nulle (sous H_1 le modèle est saturé, p.ex. si $n = p$), contrairement au test de Fisher. Ceci permet de faire des tests de H_0 contre le modèle complet dans le cas des tables de contingence du § IV.2.2.

Mentionnons également la statistique de Pearson, utilisée au même titre que la déviance et qui vaut $\sum_i (y_i - \mu_i)^2/V(\mu_i)$. Elle vaut également RSS dans le cas Gaussien. Lorsque φ est connu (modèle binomial ou poissonnien) cette statistique divisée par φ suit un χ^2_{n-p}, ce qui permet de faire un **test d'ajustement (goodness of fit test)**. Un autre test d'ajustement construit dans le même esprit, pour le modèle binomial ou poissonnien, est le test de Hosmer–Lemeshow.

Il existe des analogues du R^2 mais leur usage n'est pas recommandé, particulièrement pour les modèles logistiques ([12], § 5.2.5), car leur interprétation n'est pas immédiate, et ils sont souvent assez faibles. Le mieux est d'utiliser comme substitut une mesure de performance en prédiction plus directe, par exemple basée sur la courbe ROC présentée plus bas.

IV.4.3 Analyse de déviance

La déviance va jouer un rôle analogue au RSS de l'analyse de variance. Un exemple de table d'analyse de déviance est la table IV.1 ci-dessous.

	Df	Dev.	Resid. Dev.	Pr($>\chi$)
NULL			1025.57	
Sexe	1	228.93	796.64	0.00
Classe	2	73.05	723.59	0.00
Age	1	28.45	695.14	0.00
Sexe*Classe	2	30.30	664.84	0.00
Sexe*Age	1	14.89	649.95	1e-04
Classe*Age	2	8.58	641.37	0.01
Sexe*Classe*Age	2	1.73	639.64	0.42

TABLE IV.1 – Analyse de déviance. Les individus sont 756 passagers du Titanic pour lesquels on possède l'âge, le sexe et la classe (1re, 2e ou 3e) ; source : OzDASL. La réponse est 1 ou 0 selon que le passager a survécu ou non. On a mis un modèle binomial avec lien logit. Logiciel R.

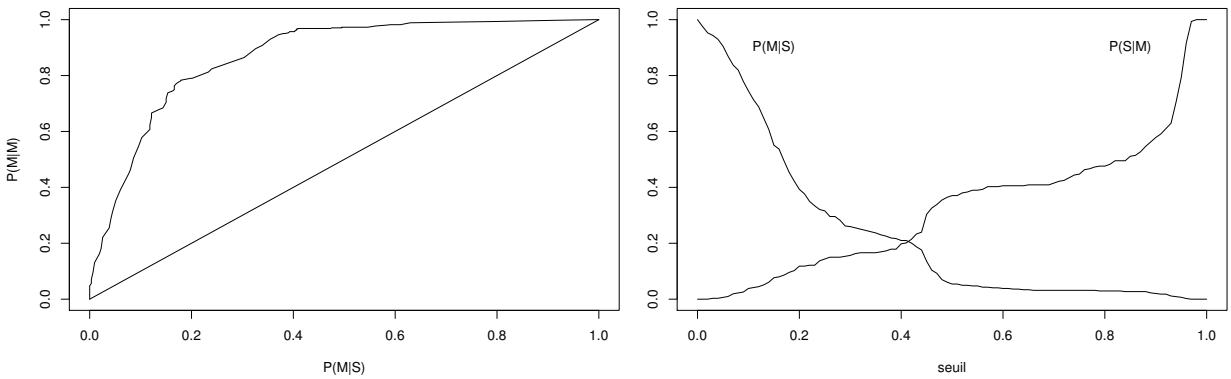
Il s'agit ici de tests emboîtés : commande `anova(glm(y~ S*C*A,,fam=binomial),test="Chi")` (ne pas oublier `test="Chi"`). La première colonne contient $p - p_0$, la deuxième est la diminution de déviance $D_0 - D$ entre deux modèles successifs, la troisième est la déviance, et la dernière le niveau (pour un $\chi^2_{p-p_0}$ sur $D_0 - D$, car $\varphi = 1$).

La fonction `step` de R fonctionne également avec ces modèles.

Courbe ROC pour le modèle logistique. Le but du modèle logistique étant souvent de faire de la prédiction (cf. l'exemple du Credit Scoring page 9) on s'intéresse à la performance de la méthode de classification $\widehat{y} = 1_{x\widehat{\beta}>\lambda}$ où λ est un seuil à choisir. L'idée est de calculer pour chaque valeur de λ la probabilité de fausse alarme (décider à tort $y = 1$) et la probabilité de bonne détection (décider correctement $y = 1$). Elles sont estimées, de préférence sur un ensemble de données n'ayant pas servi à l'estimation de β (ou par une méthode de type validation croisée), par les formules

$$PFA = \sum_i 1_{\widehat{y}_i=1,y_i=0} / \sum_i 1_{y_i=0}, \qquad PD = \sum_i 1_{\widehat{y}_i=1,y_i=1} / \sum_i 1_{y_i=1}.$$

La courbe contenant les points de coordonnées (PFA, PD) est la courbe ROC (Receiver Operating Characteristic curve). Le modèle est d'autant meilleur que la courbe longe les axes $x = 0$ puis $y = 1$. Dans notre exemple, si l'on considère maintenant l'évènement $y = 1$ comme étant la mort du passager, on trouve la courbe suivante (M=mort, S=Survie) :



On voit sur la figure de gauche (courbe ROC) l'existence d'un seuil permettant de prédire 80% des morts en ne faisant mourir à tort que 20% des survivants ; la bissectrice correspond à la performance de l'algorithme consistant à choisir au hasard 0 ou 1. La figure de droite représente l'évolution des deux risques en fonction du seuil ; on y voit que le seuil mentionné est un peu supérieur à 0,4.

Sous R, on peut utiliser la bibliothèque ROCR :

```
library(ROCR)
mod=glm(y~ ...,fam=binomial)
pred=prediction(fitted(mod),y)
```

```
perf=performance(pred, "tpr", "fpr")
plot(perf)
```

L'aire sous la courbe ROC. Il est facile de vérifier qu'elle vaut

$$A = \frac{\sum_{i,j} 1_{x_i\widehat{\beta}>x_j\widehat{\beta}} 1_{y_i=0,y_j=1}}{\sum_{i,j} 1_{y_i=0,y_j=1}} = \widehat{P}(x\widehat{\beta} < x'\widehat{\beta} \,|\, y=0, y'=1)$$

la probabilité empirique que pour deux individus de réponses distinctes pris au hasard, l'ordre obtenu sur les $x\widehat{\beta}$ soit conforme aux réponses. C'est pourquoi A est souvent pris comme mesure de qualité du modèle.

On l'obtient sous R avec cette fois

```
perf=performance(pred, "auc")
aire=perf@y.values
```

Odds ratio (rapport des cotes). Considérons le modèle additif. Soit p_f la probabilité de survie pour une femme x_f et p_h la probabilité de survie pour un homme x_h. On a en raison du lien logistique

$$\frac{p_f}{1-p_f} = \frac{\frac{1}{1+e^{-x_f\beta}}}{1-\frac{1}{1+e^{-x_f\beta}}} = e^{x_f\beta}$$

et par conséquent le rapport avec la même expression pour un homme donne

$$\frac{p_f}{1-p_f}\frac{1-p_h}{p_h} = e^{(x_f-x_h)\beta}$$

et donc si la femme et l'homme ont même âge et même classe, *cette quantité ne changera pas, quel que soit l'âge ou la classe qu'ils partagent* car le modèle est additif. Ce rapport est appelé « odds ratio » (OR) ou « rapport des cotes ». Notons que le rapport plus naturel à considérer p_f/p_h (appelé « risque relatif » RR) dépend lui des autres variables ; il est donc ici inadéquat. Il faut noter que si l'OR est malaisé à interpréter, il est fréquent que les probabilités p_h et p_f soient petites (particulièrement quand l'OR est très petit ou très grand), auquel cas OR et RR coïncident presque.

On trouve sur les données « Titanic » à partir du modèle Sexe+Classe*Age

$$\frac{p_f}{1-p_f}\frac{1-p_h}{p_h} = 14,7$$

Il valait mieux être une femme. Si l'on fait le modèle Age*Classe+Sexe*Classe, on trouve un odds ratio par classe :

$$OR_1 = 39,9 \qquad OR_2 = 76,8 \qquad OR_3 = 4,43.$$

C'est en deuxième classe que la différence entre sexes est la plus criante. On voit que l'interaction avec Classe rend tout plus compliqué, et si la Classe était une variable quantitative, on aurait un continuum d'OR ; il faut donc éviter de présenter des odds-ratio pour des variables intervenants dans des interactions, un solution pouvant être d'utiliser un modèle sans cette interaction si cette dernière influe peu, on obtient alors une sorte d'OR moyen.

Pour comparer les classes, on peut faire le modèle Age*Sexe+Classe et comparer les paires 1 et 2, puis 2 et 3 :

$$OR(1/2) = 69,7 \qquad OR(2/3) = 3,04.$$

Si une variable explicative est quantitative, par exemple l'âge, il arrive qu'on calcule l'OR associé à une variation δ de la variable entre deux individus (p.ex. δ =10 ans) :

$$\log\left(\frac{p_{a+\delta}}{1-p_{a+\delta}}.\frac{1-p_a}{p_a}\right) = \delta\beta_a.$$

Ceci peut conduire à ce genre de proposition : «Les experts ont conclu que chaque portion de 50 grammes de viande transformée consommée quotidiennement accroît le risque de cancer colorectal de 18%» [1].

On obtient directement des intervalles de confiance pour les OR à partir des ceux pour les coefficients, puisqu'il y a juste une exponentiation, d'où la commande sous R : `exp(confint(glm(...)))`.

Parenthèse : OR et RR en biostatistiques. Oublions ici les autres variables et considérons la table de contingence à quatre cases correspondant au croisement des variables h/f et s/d (survie/décès). L'objet est de proposer une mesure de l'influence de la première variable sur la réalisation de la seconde. De manière générale l'OR est souvent préféré pour les raisons suivantes (en dehors de l'avantage déjà mentionné lié à l'impossibilité d'estimer RR en présence d'autres variables ; rappelons aussi qu'il est fréquent que les probabilités p_h et p_f soient petites auquel cas OR et RR coïncident presque) :

- Si l'on remplace l'évènement « survie » par l'évènement « décès » pour le calcul du RR, on obtient $\frac{1-p_h}{1-p_f}$ qui n'est pas fonction du RR de départ, tandis que l'OR est simplement remplacé par son inverse car, avec des notations évidentes on a OR$=\frac{N_{hs}N_{fd}}{N_{hd}N_{fs}}$. Il y a donc en fait deux RR mais un seul OR.
- Lors des « études de cas témoins » (« case-control studies ») on tire d'abord au hasard un nombre équivalent de personnes guéries (ayant survécu...) et d'autres malades (décédées...) afin d'avoir suffisament d'individus dans les deux situations et ensuite on sépare chaque groupe en deux (traitement/non-traitement, classe1/classe2...). L'exemple suivant (Table 3 de [80]) concerne les accidents veineux thrombo-emboliques en Europe selon l'utilisation ou non de contraceptifs oraux où l'on a tiré au hasard 433 personnes ayant eu un accident veineux et 1044 n'en ayant pas eu

	Contraceptifs	Pas de contraceptifs	Total
Cas d'accident	265	168	433
Contrôles	356	688	1044
Total	621	856	1477

  La proportion de 433/1044 ne reflète ici aucune réalité ; on ne peut pas estimer la probabilité d'un accident pour un individu utilisant un contraceptif, qui n'a rien à voir avec 265/621, et pas davantage le RR. En revanche 265/433 est bien une estimation de la probabilité d'utiliser un contraceptif sachant que l'on a eu un accident veineux, et de même pour les trois autres rapports analogues ; par conséquent en utilisant la formule de Bayes (A=accident, C=contraceptif, $\bar{A}$=non-A), on obtient

$$OR = \frac{P(A|C)P(\bar{A}|\bar{C})}{P(\bar{A}|C)P(A|\bar{C})} = \frac{P(A,C)P(\bar{A},\bar{C})}{P(\bar{A},C)P(A,\bar{C})} = \frac{P(C|A)P(\bar{C}|\bar{A})}{P(C|\bar{A})P(\bar{C}|A)} = \frac{265\times 688}{168\times 356} \simeq 3.$$

  La probabilité d'accident étant sans doute très faible, 3 est proche du $RR = P(A|C)/P(A|\bar{C})$.

IV.5 Analyse des résidus

Les résidus standardisés sont

$$r_i = \frac{t(y_i) - t(\widehat{\mu}_i)}{t'(\widehat{\mu}_i)\sqrt{\varphi V(\widehat{\mu}_i)}\sqrt{1-h_i}}, \quad h_i = D_{ii}[\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T]_{ii} = D_{ii}\check{x}_i(\check{X}^T D\check{X})^{-1}\check{x}_i^T$$

où $t(.)$ est une certaine fonction. Si $t(x) = x$, on retrouve une formule très analogue à celle des modèles linéaires, sauf qu'il faut prendre garde à utiliser les régresseurs normalisés. Les résidus de Pearson sont simplement $(y_i - \widehat{\mu}_i)/\sqrt{V(\widehat{\mu}_i)}$.

Le but de l'introduction de t est d'avoir pour r_i une loi aussi « proche » que possible de la loi normale standard. Dans [2], il est proposé

$$t(x) = \int_0^x V(\mu)^{-1/3}\, d\mu.$$

1. Rapport de l'OMS sur la «Cancérogénicité de la consommation de viande rouge et de viande transformée». http ://www.who.int/mediacentre/news/releases/2015/cancer-red-meat/fr

En pratique, on peut soit utiliser cette formule quand on peut la calculer, soit utiliser une approximation, soit prendre $t(x) = x$ et estimer les quantiles de r_i par simulation.

Les données aberrantes seront donc détectées par les valeurs anormalement grandes des $|r_i|$. Les données isolées pourront être repérées avec les h_i. Pour le repérage des données influentes, on peut utiliser la distance de Cook

$$D_i = \frac{(y_i - \widehat{\mu}_i)^2}{\varphi V(\widehat{\mu}_i)} \ \frac{h_i}{(1-h_i)^2}.$$

On trace souvent les résidus en fonction de la réponse pour conforter l'hypothèse d'homoscédasticité, et repérer les individus aberrants. Il est difficile en pratique de faire une analyse plus fine.

V

Régression non-linéaire avec bruit additif

V.1 Modèle

On se donne le modèle pour les données :

$$y_i = f(\theta^*, x_i) + u_i, \quad u \sim \mathcal{N}(0, \sigma_*^2 Id).$$

On suppose le vecteur u gaussien pour simplifier l'exposé. θ est le paramètre à estimer. Il arrive que la variance du bruit soit également modélisée comme une fonction des variables explicatives, $E[u_i^2] = \sigma(\theta^*, x_i)^2$, avec souvent une forme qui suggère qu'elle augmente avec la moyenne, typiquement $\sigma^2 = a + b|f(\theta, x_i)|^q$. En réalité les x_i ne jouent aucun rôle et il est bien plus simple de considérer le modèle général

$$y_i = f_i(\theta^*) + u_i, \quad u \sim \mathcal{N}(0, \sigma_*^2 Id).$$

où les f_i sont des fonctions différentes connues.

On conseille les références [13, 14] pour ce qui concerne les exemples et les liens avec la pratique et [54] pour les aspects plus théoriques.

Exemple 1 : Modèle pharmaceutique monoexponentiel. On mesure l'évolution de l'efficacité d'un médicament (concentration en produit actif) au cours du temps (x_i représente le temps) :

$$y_i = \theta_1 e^{-\theta_2 x_i} + u_i.$$

L'équation physique est $y = \theta_1 e^{-\theta_2 x}$ et l'on peut tout aussi bien prendre le modèle $\log y_i = \log \theta_1 - \theta_2 x_i + u_i$ qui est linéaire, mais si u_i est i.i.d. dans un cas, il ne peut l'être dans l'autre ; il y a donc un choix à faire.

Exemple 2 : Modèles de microbiologie prévisionnelle. Il s'agit de modéliser le développement de bactéries (Listeria, etc.) au cours du temps. Soit y le nombre de bactéries et x le temps, les deux modèles classiques suivants sont le modèle de Baranyi et Roberts et le modèle de Rosso :

$$f_{BR}(\theta, x) = \theta_0 + \frac{\theta_1}{\theta_2 + e^{-\theta_3 x}}, \qquad f_R(\theta, x) = \begin{cases} \theta_0, & x < \lambda \\ \frac{\theta_1}{\theta_2 + e^{-\theta_3 x}}, & x \geqslant \lambda \end{cases}$$

Le deuxième modèle n'a que 4 paramètres en raison de la condition de continuité en $x = \lambda$.

Exemple 3 : Cinétique chimique. On considère une réaction enzymatique

$$\mathrm{E} + \mathrm{S} \underset{\mathrm{k_0}}{\overset{\mathrm{k_1}}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{\mathrm{k_2}} \mathrm{E} + \mathrm{P} \qquad (\mathrm{E} = \text{Enzyme}, \mathrm{S} = \text{Substrat}, \mathrm{P} = \text{Produit})$$

Le substrat [S] est converti en produit $[P]$.

Si [S] $\gg$ [E] et $k_0 \gg k_2$ on a la relation de Michealis-Menten qui fait intervenir la concentration initiale $[E_0]$ en E [1]

$$\frac{d[P]}{dt} = V_m \frac{[S]}{K_s + [S]}, \quad V_m = k_2[E_0], \quad K_s = \frac{k_0 + k_2}{k_1}.$$

Si l'on prend des mesures (x_i, y_i) où $y_i = d[P]/dt$ et $x_i = [S]$ on est conduit au modèle de régression

$$y_i = \frac{\theta_1 x_i}{x_i + \theta_2} + u_i.$$

Exemple 4 : Evolution d'une tumeur [27]. On postule le modèle général suivant pour l'évolution du diamètre d'une tumeur en traitement :

$$x(t) = x_0 \left(1 + k_1 t - k_2 T(1 - e^{-(t-\tau)_+/T}) - k_3(t-\tau)_+\right).$$

On suppose que le traitement a débuté à $t = 0$. Parmi les quatre termes, les deux derniers n'ont d'effet que pour $t > \tau$, et les deux premiers indiquent une vitesse d'évolution linéaire. Le paramètre τ représente l'instant initial d'une nouvelle phase au cours de laquelle la vitesse d'évolution chute d'abord à $k_1 - k_2 - k_3$ pour passer progressivement à $k_1 - k_3$. On pose $\theta = (k_1, k_2, k_3, T, \tau)$ et l'on postule le modèle suivant pour les observations y_{ij} du diamètre de la tumeur du patient i au j-ième instant de mesure t_{ij}

$$y_{ij} = x(t_{ij}, \theta_i) + e_{ij}$$

où l'on a ajouté θ_i pour indiquer que le paramètre dépend du patient. Le modèle proposé pour cette dépendance est

$$\theta_i = X_i \beta$$

où X_i est un vecteur ligne contenant les variables explicatives et β est une matrice dont la k-ième colonne permet la prédiction de la la k-ième composante de θ_i. En réalité, les auteurs désirent prendre en compte le fait que la relation ci-dessus est incomplète, et qu'il reste une partie non-expliquée, aléatoire, dans les paramètres, si bien que le modèle finalement considéré pour θ_i est

$$\theta_i = X_i \beta + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \Omega).$$

Il s'agit d'un modèle de données longitudinales à effets aléatoires. La matrice Ω donne l'ordre de grandeur de l'incertitude sur les paramètres prédits, et donne également de possibles corrélations entre eux.

V.2 Estimation des paramètres

La log-vraisemblance (du modèle à variance fixe) est $-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - f_i(\theta))^2 - n\log(\sigma)$ si bien que la méthode du maximum de vraisemblance conduit à

$$\widehat{\theta}_n = \arg\min_\theta Q(\theta), \qquad Q(\theta) = \sum_{i=1}^n (y_i - f_i(\theta))^2.$$

La solution de ce problème peut être numériquement assez difficile à trouver et ce point ne sera pas discuté ici. On peut ensuite estimer σ_*, au maximum de vraisemblance ou par validation croisée :

$$\widehat{\sigma}^2_{MV} = \frac{1}{n} Q(\widehat{\theta}), \quad \widehat{\sigma}^2_{CV} = \frac{1}{n}\sum_{i=1}^n (y_i - f_i(\theta^{(i)}))^2$$

où $\theta^{(i)}$ est obtenu sans utiliser la i-ième donnée.

1. On atteint rapidement un équilibre où $\frac{d[ES]}{dt} = 0$, ce qui conduit à $k_1[S][E] - k_0[ES] - k_2[ES] = 0$. Il ne reste plus qu'à éliminer [E] à l'aide de $[E_0] = [E] + [ES]$, puis à remplacer [ES] par son expression en fonction de $[E_0]$ et [S] dans $\frac{d[P]}{dt} = k_2[ES]$.

V.3 Utilisation du bootstrap et du Monte-Carlo

On démontre mathématiquement des propriétés de convergence, mais elles sont de nature asymptotique, et leur validité pour n fini peut dépendre très fortement de chaque cas particulier. Il conviendra de vérifier le bon fonctionnement des algorithmes utilisés par des simulations à (θ^*, σ_*) connus ; elles permettront de

1. Vérifier les propriétés de convergence et la validité des algorithmes
2. Estimer la matrice de covariance de $\widehat{\theta}$ et fournir des intervalles de confiance.

On peut faire ces simulations de différentes façons :

1. Simulation (ou « bootstrap paramétrique »). Générer des données avec des (θ, σ) différents, et éventuellement les x_i de l'expérience, ceci S fois (p.ex. $S = 10000$) :

$$y_i^s = f_i(\theta) + u_i^s, \quad u^s \sim \mathcal{N}(0, \sigma^2 Id), \quad i = 1, \dots n, \quad s = 1, \dots S.$$

   Les estimées $\widehat{\theta}^s$ permettent de vérifier le bon comportement de l'algorithme, et même d'obtenir, par exemple, une estimation de la variance d'estimation de $\widehat{\theta}$ sous la loi (θ, σ) :

$$Var(\theta, \sigma) \simeq \frac{1}{S} \sum_{s=1}^{S} (\widehat{\theta}^s - \theta)(\widehat{\theta}^s - \theta)^T.$$

   Le choix $\theta = \widehat{\theta}$ et $\sigma = \widehat{\sigma}$ dans cette expérience conduit à une estimée de $Var(\widehat{\theta}, \widehat{\sigma})$ qui est (on peut l'espérer) un bon estimateur de $Var(\theta^*, \sigma_*)$, variance de $\widehat{\theta}$.
   De même, soit θ (a priori proche de θ^*) et un (δ, δ') tel que l'intervalle $I = [\theta - \delta, \theta + \delta']$ contienne 95% des $\widehat{\theta}^s$; alors $\theta \in [\widehat{\theta}^s - \delta', \widehat{\theta}^s + \delta]$ pour 95% des valeurs de s, et $[\widehat{\theta}^s - \delta', \widehat{\theta}^s + \delta]$ est donc un intervalle de confiance de niveau approximativement égal à 5%. Ce qui conduit à utiliser $[\widehat{\theta} - \delta', \widehat{\theta} + \delta]$ comme intervalle de confiance ; cette démarche présuppose que la statistique $\widehat{\theta} - \theta^*$ est (localement) pivotale car la paire (δ, δ') n'a pas été calculée avec θ^* (inconnu) mais avec un autre θ supposé proche (en pratique $\widehat{\theta}$). Sinon il faut en toute rigueur prendre une paire (δ, δ') qui convienne pour toute valeur de θ.

2. Bootstrap sur les résidus. Pour estimer la loi de $\widehat{\theta}$, on simulait ci-dessus de nouvelles observations avec la loi donnée par $(\widehat{\theta}, \widehat{\sigma})$. On se propose ici de modifier la simulation des résidus : on génère de nouveaux y_i avec les x_i de l'expérience et $\widehat{\theta}$

$$y_i^b = f_i(\widehat{\theta}) + u_i^b, \quad b = 1, \dots B$$

   où chaque u_i^b sera tiré indépendamment à partir d'une loi uniforme sur l'ensemble $\{\tilde{u}_1, \dots \tilde{u}_n\}$, la suite $\tilde{u}$ étant formée des $\widehat{u}_i = y_i - f_i(\widehat{\theta})$ (ou mieux $y_i - f_i(\widehat{\theta}^{(i)})$) empiriquement recentrés [43].
   On dispose donc maintenant de B suites de données et B estimateurs $\widehat{\theta}^b$, dont la variance empirique donne une estimée de la variance de $\widehat{\theta}$.
   Variante : Le wild bootstrap est recommandé si les résidus ne sont pas i.i.d. Il évite de mélanger des résidus entre individus. Il s'agit de prendre $u_i^b = z_i \widehat{u}_i$ où les z_i sont tirés avec la loi suivante [58, 61]

$$z_i = \begin{cases} (1 - \sqrt{5})/2 & \text{avec probabilité} \quad (1 + \sqrt{5})/(2\sqrt{5}) \\ (1 + \sqrt{5})/2 & \text{avec probabilité} \quad (-1 + \sqrt{5})/(2\sqrt{5}). \end{cases}$$

   Cette variable satisfait $E[z] = 0, E[z^2] = E[z^3] = 1$ (noter que z_i est racine de $x^2 - x - 1 = 0$). On a donc préservé les moments jusqu'à l'ordre trois.

V.4 Propriétés asymptotiques

On s'intéresse au comportement de l'estimateur quand le nombre de données n augmente. Sous les hypothèses habituelles dans le cadre du maximum de vraisemblance, on aura la convergence presque sûre

de $\widehat{\theta}_n$ vers θ^* avec la normalité asymptotique quand $n \to \infty$

$$\Gamma_n^{1/2}(\widehat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, \sigma_*^2 Id)$$

$$\Gamma_n = \check{X}^T \check{X}, \quad \check{X} = \begin{pmatrix} \check{x}_1 \\ \vdots \\ \check{x}_n \end{pmatrix}, \quad \check{x}_i = \nabla_\theta f_i(\theta^*)$$

$\nabla_\theta f_i(\theta^*)$ est ici un vecteur ligne. On peut remplacer θ^* par $\widehat{\theta}_n$ dans la calcul de Γ_n, auquel cas on notera cette matrice $\widehat{\Gamma}_n$.

V.5 Régions de confiance

V.5.1 Régions théoriques

Des résultats de l'annexe C, on déduit les régions de confiance (asymptotiques) de niveau α après remplacement de σ_* par $\widehat{\sigma}$

$$\mathcal{R}_\alpha(Y) = \left\{ \theta : \frac{Q(\theta) - Q(\widehat{\theta})}{\widehat{\sigma}^2} \leqslant \chi_p^2(1-\alpha) \right\} \quad \text{ou} \quad \left\{ \theta : \frac{1}{\widehat{\sigma}^2}(\theta - \widehat{\theta})\widehat{\Gamma}(\theta - \widehat{\theta}) \leqslant \chi_p^2(1-\alpha) \right\}.$$

où Y désigne l'ensemble des données.

V.5.2 Ajustement du niveau par simulation ou bootstrap

La région de confiance $\mathcal{R}_\alpha(Y)$ a un niveau réel α' différent de α. On peut l'estimer à partir des données bootstrappées (sous $\widehat{\theta}$) Y^b : α' sera donné par la proportion de b tels que $\widehat{\theta} \notin \mathcal{R}_\alpha(Y^b)$.

V.5.3 Intervalles de confiance

En appliquant les résultats du C.2 avec $g(\theta) = \theta_j$, on obtient l'intervalle de confiance

$$I_\alpha(Y) = [\widehat{\theta}_{nj} - \delta, \widehat{\theta}_{nj} + \delta], \qquad \delta = \widehat{\sigma}[(\widehat{\Gamma}_n)^{-1}]_{jj}^{1/2} t_{n-1}(1-\alpha/2)$$

où l'on a remplacé la racine d'un χ^2 par un Student pour avoir une formule analogue au cas linéaire, cf. III.2.1, ce qui ne change rien dans le cadre asymptotique $n \to \infty$.

V.6 Tests

De la même façon, on a les tests classiques de l'annexe C. Par exemple, le test du maximum de vraisemblance pour $g(\theta) = 0$ s'écrit « $n \log \frac{Q(\widehat{\theta}_n^0)}{Q(\widehat{\theta}_n)} \leqslant \chi_q^2(1-\alpha)$ » où $\widehat{\theta}_n^0$ est l'estimée au maximum de vraisemblance sous la contrainte $g(\theta) = 0$, et q est la dimension de g.

Aspects pratiques. Pour n petit, le seuil $\chi_q^2(1-\alpha)$ est une mauvaise approximation du seuil réel. Il sera bon de réévaluer le quantile en faisant des simulations du membre de gauche sous H_0. Notons pour un ensemble de données Y, $T(Y) = \log \frac{Q(\widehat{\theta}_n^0)}{Q(\widehat{\theta}_n)}$, alors on pourra employer la méthode suivante valide pour toute statistique de test $T(Y)$:

1. Estimer $\widehat{\theta}_n^0(Y)$
2. Simuler des ensembles des données Y^s (ou Y^b) comme au § V.3 sous la loi associée à $\widehat{\theta}_n^0(Y)$
3. Calculer les $T(Y^s)$
4. le seuil sera la valeur λ telle qu'une proportion α seulement des $T(Y^s)$, $s = 1, \dots S$ dépassent cette valeur.
5. la p-value associée à $T(Y)$ sera la proportion de s tels que $T(Y^s) > T(Y)$.

V.7 Analyse des résidus

En linéarisant le modèle au voisinage de θ^*, on obtient par des procédés standard l'approximation du residu standardisé

$$r_i = \frac{y_i - f_i(\widehat{\theta})}{\widehat{\sigma}\sqrt{1-h_i}} \qquad h_i = \left[\check{X}(\check{X}^T\check{X})^{-1}\check{X}^T\right]_{ii}.$$

Pour la détection de données influentes, on a la statistique de - :

$$C_i = \frac{h_i}{p(1-h_i)}\, r_i^2.$$

Ces statistiques sont des indicateurs qui permettent de détecter des individus particuliers ; ils sont basés sur une linéarisation qui peut être très approximative pour des n petits.

A

Sélection de modèles

La situation est la suivante : on se donne plusieurs modèles qu'on identifie et l'on veut choisir le meilleur, et par exemple savoir si un modèle compliqué est justifié. Si ce choix est motivé par un besoin de faire de la *prédiction*, les solutions que l'on va voir dans la suite sont généralement bonnes. Si au contraire il s'agit de faire de l'*interprétation* (p.ex. savoir si telle ou telle variable importe, savoir si le modèle est linéaire ou pas), c'est beaucoup plus difficile, particulièrement si l'on a à choisir parmi un nombre infini de modèles. Par exemple, un modèle non-linéaire identifié peut avoir des perfomances statistiquement tout-à-fait raisonnables même si le vrai modèle est linéaire ; en ce cas le modèle non-linéaire sera bon en prédiction mais l'interprétation juste est la linéarité. De même on peut se permettre, en prédiction, de prendre trop de régresseurs en compte, du moment que le modèle estimé leur donne un poids suffisamment faible ; cette option a de plus l'avantage de conduire à un estimateur peu biaisé ce qui facilite la construction d'intervalles de confiance[1]. Cette marge de manœuvre rend le problème de la sélection pour la prédiction plus simple. Pour des compléments concernant cette partie, nous recommandons [17].

On a déjà vu une méthode de sélection pour l'interprétation permettant de décider entre deux *modèles emboîtés*, c'est le test de Fisher. Il se généralise en (cf. § C.3.1) :

$$\boxed{\text{Rejeter } H_0 \text{ si} \quad 2(\mathcal{L}_1(y) - \mathcal{L}_0(y)) \geqslant \chi^2_q(1-\alpha)}$$

où $\mathcal{L}_i(y)$ est la log-vraisemblance de $y = (y_1, ..y_n)$ sous H_i et $q = p_1 - p_0$ est la différence entre le nombre de paramètres sous chaque hypothèse. Le principe du test de Fisher est de ne refuser l'hypothèse simple H_0 qu'en cas de valeur extrême de la statistique, et il est paramétré par α.

On présente ici des méthodes plus générales qui conviennent pour des modèles non-emboîtés, dans un cadre non-linéaire, et qui n'utilisent pas de seuil.

Appelons p le nombre de paramètres ; il est clair que le modèle le plus compliqué (p grand) aura généralement l'erreur de prédiction la plus faible. Plusieurs critères ont étés proposés pour les modèles de régression, ils pénalisent les p grands à erreur de prédiction $\sum \widehat{u}_i^2$ constante :

- Validation croisée par leave-one-out : $CV = \frac{1}{n}\sum_{i=1}^n \widehat{u}_i^2/(1-h_i)^2$ (cf. exercice 10 p. 22)
- Critère d'Akaike : $AIC = n\log(RSS) + 2p$ (si σ est connu : $AIC = \sigma^{-2}RSS + 2p$, qui coïncide avec C_p-Mallows).
- Critère de Wallace-Boulton-Schwarz [72, 48] : $BIC = n\log(\widehat{\sigma}^2) + p\log(n)$

Extension à des modèles généraux. Utilisation pratique. Ces critères s'utilisent pour des modèles paramétriques généraux, à condition de les exprimer en fonction de la log-vraisemblance des observations $y = (y_1, \dots y_n)$ (cf. exercice 1 p. 45 : $\mathcal{L}(y) = -\frac{n}{2}\log(2\pi e\widehat{\sigma}^2_{MV})$) :

1. Car c'est le biais qui est difficile à estimer. Voir p.ex. [46].

$$\boxed{\begin{array}{l} CV' = -2\sum_i \mathcal{L}(y_i/y^{(i)}) \\ AIC = -2\mathcal{L}(y) + 2p \\ BIC = -2\mathcal{L}(y) + p\log(n) \end{array}}$$

où, dans CV', chaque terme est la log-vraisemblance du i-ième échantillon quand l'estimation a été faite en utilisant les autres (CV plus haut n'est pas exactement une somme de log-vraisemblances du cas gaussien mais simplement d'erreurs de prédiction). On cherchera le modèle qui minimise la valeur du critère considéré.

L'utilisation des critères CV ou AIC peut conduire à une légère surestimation de p ; ils donnent toutefois de relativement bons résultats en *prédiction*. Si en revanche on veut faire de l'*interprétation*, BIC sera souvent meilleur car il a plutôt tendance à sous-estimer p. On trouvera les détails mathématiques dans [1].

Leave-one-out ou V-fold CV. La V-fold CV consiste à séparer les données en V sous-ensembles, à en retirer un de l'apprentissage pour l'utiliser ensuite au calcul d'erreur de prédiction ; ceci est fait successivement en extrayant chacun des V sous-ensembles ; on moyenne ensuite les V erreurs de prédictions moyennes obtenues. Le leave-one-out est donc la n-fold CV. Dans la V-fold CV, on peut préférer extraire aléatoirement un ensemble de n/V individus à chaque étape. On choisit souvent [2] $V \simeq 5$. La V-fold CV peut conduire à moins de fluctuations (cf les figures § II.9.1) pour la raison suivante : Dans le leave one out, à ordre fixe, c'est toujours en gros le même modèle qui est estimé (on ne change que deux échantillons), mais il peut y avoir des changement importants au passage d'un ordre à l'autre (instabilité du modèle). Dans le V-fold, le modèle estimé variera davantage, ce qui entraîne une meilleure moyennisation et moins de variabilité d'un ordre à l'autre ; il est très généralement préféré [73, 32]. Ceci n'empêche que l'estimée d'erreur du leave one out est plutôt bonne [73]. Pour limiter ces fluctuations, il est préférable d'utiliser les mêmes sous-ensembles pour tous les modèles à comparer, i.e. la boucle *modèles* est à l'intérieur de la boucle *splitting*, ce qui généralement complique la programmation.

> **La « 1-se-rule » pour la CV.** Dans la figure II.5 on voit des oscillations ou un grand plat qui posent une incertitude sur le choix du minimum de la courbe. La 1-se-rule consiste à se placer en ce point, calculer l'écart-type s de sa valeur (c'est une moyenne empirique) et à chercher le premier (en partant des modèles les plus simples) pour lequel la valeur est à moins de s du premier, cf. [11] p. 244 ou [34] § 3.4.3.

Interprétation de CV et AIC. Ce sont deux estimateurs différents de la la qualité d'ajustement mesurée sur un échantillon extérieur à l'estimation. En régresssion linéaire $\exp(AIC/n) = \widehat{\sigma}^2 e^{p/n} \simeq \widehat{\sigma}^2(1-p/n)^{-1}$ qui est CV dans le cas où les h_i sont constants (égaux à p/n). AIC n'est valide que si l'estimation a été faite au maximum de vraisemblance.

Interprétation de BIC (MDL). Si l'on cherche à coder les réponses pour les transmettre à quelqu'un qui possède déjà les régresseurs, la méthode la plus économique consiste à transmettre le paramètre du modèle estimé et les erreurs de prédiction du modèle, avec une certaine précision correspondant à celle requise pour les réponses (pour une précision requise inférieure à $\widehat{\sigma}$, on ne transmet donc que le paramètre). Wallace et Boulton ont remarqué en 1968 que le nombre de bits nécessaires à cette opération vaut en première approximation $\frac{1}{2}BIC - n\log(\varepsilon)$ où ε est la précision requise sur les réponses. Le terme $p\log(n)/2$ est le coût de la transmission des paramètres avec une précision adéquate (qui est d'ordre $1/\sqrt{n}$). On voit donc que minimiser BIC consiste à choisir le modèle le plus économique pour la transmission des réponses. Ce critère est appelé aussi MDL (minimum description length). BIC peut aussi se justifier par une approche bayésienne générale due à Gideon Schwarz [72].

2. Sur CV et ses variantes, voir [36]. Voir aussi la discussion de [31].

B

Régression PLS

Nous renvoyons au § II.9 et § II.9.1 pour les exemples et la motivation de la méthode. Mentionnons la référence [21] qui décrit la méthode et ses variantes et le § 3.5.2. de [11].

Considérons d'abord le cas où il n'y a qu'une seule réponse. La méthode PLS consiste à extraire une famille de variables dites « latentes », de la forme Xw avec $\|w\| = 1$, orthogonales entre elles, de sorte à maximiser la somme des carrés des covariances de ces dernières avec y. Il se trouve que, comme pour l'ACP, lorsque l'on fait croître cette famille, on ne fait que rajouter des variables sans avoir à remettre en question les précédentes, ce qui fait que l'on définit bien une suite de variables.

Noter que si au lieu de la covariance on maximise la corrélation, sans la contrainte $\|w\| = 1$ qui devient inutile, la première variable latente que l'on récupère est le $\widehat{y} = X\widehat{\beta}$ de la régression linéaire habituelle et l'algorithme s'arrête, car toute variable de la forme Xw orthogonale à $X\widehat{\beta}$ est orthogonale à y.

Dans le cas de réponses multiples, chaque y_i est un vecteur ligne, si bien qu'on a une matrice Y et un espace vectoriel $\mathcal{Y}$ engendré par les colonnes de Y. La méthode de régression linéaire fonctionne comme avant (cf. § II.8) avec $\widehat{\beta} = (X^T X)^{-1} X^T Y, \widehat{Y} = X\widehat{\beta}$, et $\widehat{\beta}$ est la matrice qui contient les résultats des régressions linéaires faites séparément pour chaque colonne de Y. La méthode PLS quant à elle fera apparaître des vecteurs de $\mathcal{X}$ et $\mathcal{Y}$ de plus grande covariance maximale.

L'algorithme (cf. [21] p.141). Il consiste à calculer la paire de vecteurs $x \in \mathcal{X}$ et $y \in \mathcal{Y}$ de plus grande covariance (sous une contrainte particulière) puis à orthogonaliser $\mathcal{X}$ à x, et à recommencer ; on retire donc à chaque fois à chaque colonne de X sa prédiction par x, cette matrice sera notée X_a dans la suite, $X_0 = X$.

L'algorithme est [1], avec $X_0 = X$, $a = 1, 2, \ldots$:

$$(w_a, u_a) = \arg\max_{u,w} \{\langle X_{a-1} w, Yu\rangle : \|u\| = 1, \|w\| = 1\}$$

$$t_a = X_{a-1} w_a / \|X_{a-1} w_a\| \quad \text{(nouvelle composante orthogonale)}$$

$$X_a = X_{a-1} -_a (t_a^T X_{a-1}) \quad \text{(orthtogonalisation des colonnes à } t_a\text{)}$$

noter que l'orthogonalisation des colonnes de X peut se faire pas à pas car les t_a sont orthogonaux. La résolution en (w, u) donne pour w le vecteur propre de $X_{a-1}^T Y Y^T X_{a-1}$ associé à la valeur propre maximale (vecteur singulier maximal à droite de $Y^T X_{a-1}$; si Y est un vecteur c'est $X_{a-1}^T Y$) [2].

Comme $\{t_1, \ldots t_a\}$ et $\{Xw_1, \ldots Xw_a\}$ engendrent le même espace (vérifier !), on choisira ces derniers comme nouveaux régresseurs.

1. De même que pour l'ACP, les colonnes de X seront généralement centrées ce qui fait que les produits scalaires apparaissant dans la suite sont des covariances empiriques, mais ce n'est pas absolument nécessaire.

2. Noter au passage l'orthogonalité des w_a : Pour le vérifier, remarquer que $X_a w_a = 0$, en déduire par récurrence que $X_b W_a = 0$, $b \geqslant a$; par conséquent, si $b > a$, w_a est dans le noyau de $X_{b-1}^T Y Y^T X_{b-1}$ et donc orthogonal à X_b.

Un autre point de vue. On peut très bien réécrire le problème d'optimisation de manière équivalente :

$$(w_a, u_a) = \arg\max_{u,w}\{\langle Xw, Y_{a-1}u\rangle : \|u\| = 1, \|w\| = 1\}$$

où $Y_a = (1 - T_a T_a^T)Y$ et $T_a = [t_1, \dots t_a]$, car $X_a = (1 - T_a T_a^T)X$. On cherche à chaque étape le vecteur Xw de $\mathcal{X}$ de plus grande covariance avec les résidus de prédiction des réponses basée sur les composantes précédentes, sous la contrainte $\|w\| = 1$.

Les axes principaux en réponse. En posant $W_a = [w_1, \dots w_a]$, on obtient la régression :

$$\begin{aligned} X'_a &= XW_a \\ \widehat{\beta}_a &= ({X'_a}^T X'_a)^{-1} {X'_a}^T Y \\ \widehat{Y}_a &= X'_a \widehat{\beta}_a. \end{aligned}$$

Si l'on retient moins de composantes que la dimension de $\mathcal{Y}$, chaque ligne de $\widehat{Y}_a$ sera combinaison linéaire des lignes de $\widehat{\beta}_a$, les « axes principaux en réponse ».

C

Asymptotique du maximum de vraisemblance

C.1 Théorèmes-limite

On se donne une famille de lois $P_{\theta,x}$ dépendant d'un paramètre $\theta \in \mathbb{R}^d$ et d'un régresseur x ; elles possèdent une densité $p_{\theta,x}(y)$ par rapport à une mesure commune $\nu_x(dy)$. On observe une suite de variables aléatoires $(y_i)_{i=1,\dots n}$ indépendantes de loi P_{θ^*,x_i}.

L'estimateur au maximum de vraisemblance de θ^*, la vraisemblance est

$$\widehat{\theta}_n = \max_\theta \mathcal{L}_n(\theta)$$
$$\mathcal{L}_n(\theta) = \sum_i \log p_{\theta,x_i}(y_i).$$

On désignera par $\mathcal{L}'$ et $\mathcal{L}''$ les dérivés premières (vecteur) et seconde (matrice) de la fonction $\mathcal{L}(\theta)$. Les résultats qui suivent s'obtiennent heuristiquement sans difficulté, les preuves rigoureuses sont en revanche délicates. La matrice d'information de Fisher est définie par

$$I_n(\theta) = -E_\theta\left[\mathcal{L}''_n(\theta)\right] = E_\theta\left[\mathcal{L}'_n(\theta)\mathcal{L}'_n(\theta)^T\right].$$

Sous certaines hypothèses que nous ne détaillerons pas, et qui ont essentiellement trait d'une part à la régularité en θ de la fonction $p_{\theta,x}(y)$ et d'autre part au fait que la suite $\widehat{\theta}_n$ reste bornée, et en supposant de plus que

$$\boxed{\text{Hypothèse : } I_n(\theta^*)^{-1} \longrightarrow 0}$$

(typiquement $I_n(\theta^*)$ est d'ordre n) on obtient la **convergence presque sûre** de $\widehat{\theta}_n$ vers θ^* quand n tend vers l'infini. L'hypothèse est en défaut lorsque la loi des données ne dépend pas (ou pas assez) de θ, ce qui implique bien entendu que θ^* ne peut pas être estimée à partir de ces dernières.

La propriété de loi des grands nombres suivante (somme de variables indépendantes)

$$-I_n(\theta)^{-1}\mathcal{L}''_n(\theta) \longrightarrow Id.$$

permet d'avoir aussi les estimateurs simples suivants de $I_n(\theta^*)$:

$$\boxed{I_n(\theta^*) \simeq -\mathcal{L}''_n(\widehat{\theta}_n) \simeq I_n(\widehat{\theta}_n)}$$

valides sous des hypothèses de régularité raisonnables et couramment utilisés ; dans la suite, I_n désignera $I_n(\theta^*)$ ou un estimateur consistant cette matrice.

On montre ensuite la **normalité asymptotique des scores**

$$\boxed{I_n^{1/2}\mathcal{L}_n'(\theta^*) \longrightarrow \mathcal{N}(0, Id)}$$

C'est une simple conséquence du théorème-limite central. En écrivant la dérivée de la log-vraisemblance au voisinage de θ^* il vient

$$0 = \mathcal{L}_n'(\widehat{\theta}_n) \simeq \mathcal{L}_n'(\theta^*) + (\widehat{\theta}_n - \theta^*)\mathcal{L}_n''(\theta^*)$$

soit

$$\mathcal{L}_n''(\theta^*)(\widehat{\theta}_n - \theta^*) \simeq -\mathcal{L}_n'(\theta^*)$$

et l'on montre alors la **normalité asymptotique de l'estimateur**

$$\boxed{I_n^{1/2}(\widehat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, Id)}$$

On en déduit également, en développant $\mathcal{L}_n$ au voisinage de $\widehat{\theta}_n$, la convergence en loi de la **déviance** vers un χ^2_p

$$\boxed{2\big(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta^*)\big) \longrightarrow \chi^2_p.}$$

Normalité des fonctions de l'estimateur. Supposons que I_n/n converge vers une matrice I. Soit g une fonction à valeurs dans $\mathbb{R}^q$, en développant au voisinage de $\widehat{\theta}_n$:

$$g(\widehat{\theta}_n) - g(\theta^*) = \nabla g(\widehat{\theta}_n')(\widehat{\theta}_n - \theta^*)$$

d'où

$$\boxed{\sqrt{n}\big(g(\widehat{\theta}_n) - g(\theta^*)\big) \longrightarrow \mathcal{N}(0, G^T I^{-1} G),} \qquad G = \nabla g(\theta_*).$$

Dans la suite on supposera que $q \leqslant p$, que la dérivée de g est de rang plein au voisinage de θ^* et l'on notera :

$$I_n^g = \big[G_n^T I_n^{-1} G_n\big]^{-1}, \qquad G_n = \nabla g(\widehat{\theta}_n).$$

C.2 Régions de confiance

Des résultats précédents, on déduit aussitôt les régions de confiance asymptotiques de niveau α

$$\boxed{\mathcal{R}_\alpha = \{\theta : 2(\mathcal{L}_n(\theta) - \mathcal{L}_n(\widehat{\theta}_n)) \leqslant \chi^2_p(1-\alpha)\}}$$

et

$$\boxed{\mathcal{R}_\alpha = \{\theta : \mathcal{L}_n'(\theta) I_n \mathcal{L}_n'(\theta) \leqslant \chi^2_p(1-\alpha)\}}$$

et pour les fonctions (en particulier $g(\theta) = \theta$)

$$\boxed{\mathcal{R}_\alpha = \{v : (g(\widehat{\theta}_n) - v)^T I_n^g (g(\widehat{\theta}_n) - v) \leqslant \chi^2_q(1-\alpha)\}}$$

C.3 Tests

On veut tester l'hypothèse générale

$$H_0 : \ g(\theta^*) = 0$$

pour une certaine fonction g à valeurs dans $\mathbb{R}^q$ et un niveau $1-\alpha$.

C.3.1 Test du rapport de vraisemblance

Soit $\widehat{\theta}_{0n}$ l'estimateur au maximum de vraisemblance de θ sous la contrainte $g(\theta) = 0$. On peut vérifier que si $g(\theta^*) = 0$

$$I_n^{1/2}(\widehat{\theta}_{0n} - \theta^*) = P I_n^{1/2}(\widehat{\theta}_n - \theta^*) + O(\|\widehat{\theta}_n - \theta^*\|^2)$$

où $P = Id - I_n^{-1/2} g_n' I_n^g g_n' I_n^{-1/2}$ est un projecteur orthogonal de rang q. On montre alors facilement que sous H_0

$$2(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\widehat{\theta}_{0n})) \longrightarrow \chi^2_q.$$

D'où le test :

$$\boxed{\text{Rejeter } H_0 \text{ si} \quad 2(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\widehat{\theta}_{0n})) \geqslant \chi^2_q(1-\alpha)}$$

où $\chi^2_q(.)$ désigne la fonction quantile du χ^2_q.

C.3.2 Test des scores

On montre de manière analogue que sous H_0

$$\mathcal{L}_n'(\widehat{\theta}_{0n}) I_n \mathcal{L}_n'(\widehat{\theta}_{0n}) \longrightarrow \chi^2_q.$$

D'où le test :

$$\boxed{\text{Rejeter } H_0 \text{ si} \quad \mathcal{L}_n'(\widehat{\theta}_{0n}) I_n \mathcal{L}_n'(\widehat{\theta}_{0n}) \geqslant \chi^2_q(1-\alpha).}$$

Le plus simple sera ici de prendre $I_n = -\mathcal{L}_n''(\widehat{\theta}_{0n})$.

C.3.3 Test de Wald

Il se déduit de la normalité de $g(\widehat{\theta}_n)$:

$$\boxed{\text{Rejeter } H_0 \text{ si} \quad g(\widehat{\theta}_n)^T I_n^g g(\widehat{\theta}_n) \geqslant \chi^2_q(1-\alpha).}$$

En particulier, pour tester $H_0 : R\theta^* = l$, on a :

$$\boxed{\text{Rejeter } H_0 \quad \text{si} \quad (R\widehat{\theta} - l)^T (R\mathcal{I}_n^{-1} R^T)^{-1} (R\widehat{\theta} - l) \geqslant \chi^2_q(1-\alpha).}$$

C.3.4 Aspects pratiques.

Pour n petit, le seuil $\chi^2_q(1-\alpha)$ est une mauvaise approximation du seuil réel. Il sera bon de réévaluer le quantile en faisant des simulations du membre de gauche (ou en utilisant le bootstrap, cf. chapitre V). Notons pour un ensemble de données Y, $T(Y) = 2(\mathcal{L}_n(\widehat{\theta}_n(Y)) - \mathcal{L}_n(\widehat{\theta}_{0n}(Y)))$, alors on pourra :

1. Estimer $\widehat{\theta}_{0n}(Y)$
2. Simuler des ensembles de données Y^s comme au § V.3 sous la loi associée à $\widehat{\theta}_{0n}(Y)$
3. Calculer les $T(Y^s)$
4. α sera la proportion de s tels que $T(Y^s) > T(Y)$.

Bibliographie

[1] J.-M. Azaïs et J.-M. Bardet. *Le modèle linéaire par l'exemple*. Dunod, 2005.

[2] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, 1978.

[3] L. Breiman et J. H. Friedman. "Predicting multivariate responses in multiple linear regression". In : *J. Roy. Statist. Soc. Ser. B* 59.1 (1997), p. 3-54.

[4] P. Bühlmann et S. van de Geer. *Statistics for high-dimensional data*. Springer, 2011.

[5] N. V. Chawla, K. W. Bowyer et L. O. Hall. "SMOTE : Synthetic Minority Over-sampling Technique". In : *J. Artif. Intell. Res.* 16 (2002), p. 321-357.

[6] P. J. Diggle et P. J. Ribeiro Jr. *Model-based geostatistics*. Springer, 2007.

[7] J. Fan et J. Lv. "A selective overview of variable selection in high dimensional feature space". In : *Statist. Sinica* 20.1 (2010), p. 101-148.

[8] W. A. Fuller. *Measurement error models*. Wiley, 1987.

[9] X. Guyon. *Statistique et économétrie*. Ellipses, 1991.

[10] W. Härdle. *Applied nonparametric regression*. Cambridge University Press, 1990.

[11] T. Hastie, R. Tibshirani et J. Friedman. *The elements of statistical learning*. Data mining, inference, and prediction. Springer, 2009.

[12] D. Hosmer, S. Lemeshow et R. Sturdivant. *Applied Logistic Regression*. 3e éd. Wiley, 2013.

[13] S. Huet, E. Jolivet et A. Messéan. *La régression non-linéaire, méthodes et applications en biologie*. Inra, 1992.

[14] S. Huet et al. *Statistical tools for nonlinear regression*. Springer, 1996.

[15] P. McCullagh et J. A. Nelder. *Generalized linear models*. Chapman & Hall, 1983.

[16] J. O. Ramsay et B. W. Silverman. *Functional data analysis*. Springer, 2005.

[17] C. R. Rao et Y. Wu. "On model selection". In : *Model selection*. Inst. Math. Statist., 2001, p. 1-64.

[18] S. R. Searle, G. Casella et C. E. McCulloch. *Variance components*. John Wiley, 1992.

[19] L. A. Stefanski. "Measurement error models". In : *J. Amer. Statist. Assoc.* 95.452 (2000), p. 1353-1358.

[20] M. L. Stein. *Interpolation of spatial data*. Springer, 1999.

[21] M. Tenenhaus. *La régression PLS*. Technip, 1998.

[22] V. Venables. "Exegeses on Linear Models". In : *S-PLUS User's Conference* (1998).

[23] R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Elsevier, 2012.

Références

[24] Aiji et al. "Apport du boostrap à la régression PLS". In : *Oil & Gaz Science Technology Rev. IFP* 58.5 (2003), p. 599-608.

[25] J. Anderson. "Gender-related differences on open and closed assessment tasks". In : *International Journal of Mathematics Education in Science and Technology* 33.4 (2002), p. 495-503.

[26] P. Bastien. "Modèle Cox-PLS : application en transcriptonique". In : *36è Jounées de la SFDS* (2004).

[27] Bastogne. "Phenomenological modeling of tumor diameter growth based on a mixed effects model". In : *Journal of Theoretical Biology* 262 (2010), 544-–552.

[28] D. Bates. "lme4 : Mixed-effects modeling with R". `lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`. 2010.

[29] P. J. Bickel et K. A. Doksum. "An analysis of transformations revisited". In : *J. Amer. Statist. Assoc.* 76.374 (1981), p. 296-311.

[30] F. Breidt. "Ecological Modeling with Soils Data : Semiparametric Stochastic Mixed Models for Increment Averages". In : *Journées Statistiques de Rennes* (2006).

[31] L. Breiman et P. Spector. "Submodel Selection and Evaluation in Regression. The X-Random Case". In : *International Statistical Review* 60.3 (1992), p. 291-319.

[32] L. Breiman. "Better subset regression using the nonnegative garrote". In : *Technometrics* 37.4 (1995), p. 373-384.

[33] L. Breiman. "The little bootstrap and other methods for dimensionality selection in regression : X-fixed prediction error". In : *J. Amer. Statist. Assoc.* 87.419 (1992), p. 738-754.

[34] L. Breiman et al. *Classification and regression trees*. Wadsworth Advanced Books et Software, 1984.

[35] T. S. Breusch et A. R. Pagan. "A simple test for heteroscedasticity and random coefficient variation". In : *Econometrica* 47.5 (1979), p. 1287-1294.

[36] P. Burman. "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods". In : *Biometrika* 76.3 (1989), p. 503-514.

[37] D. Causeur et F. Husson. "A 2-dimensional extension of the Bradley-Terry model for paired comparisons". In : *J. Statist. Plann. Inference* 135.2 (2005), p. 245-259.

[38] D. Cobb. "A theory of production". In : *American Economic Review* (1928), p. 139-165.

[39] J. Durbin et G. S. Watson. "Testing for serial correlation in least squares regession. III". In : *Biometrika* 58 (1971), p. 1-19.

[40] M. Eysenck. "Age differences in incidental learning". In : *Developmental Psychology* 10 (1974). Données et description précise disponibles sur le site OzDASL : `http://www.statsci.org/data/general/eysenck.html`.

[41] J. Fan et J. Lv. "Sure independence screening for ultrahigh dimensional feature space". In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70.5 (2008).

[42] R. A. Fisher. *Statistical Methods for Research Workers*. 5^{e} éd. Oliver et Boyd, 1934.

[43] D. A. Freedman. "Bootstrapping regression models". In : *Ann. Statist.* 9.6 (1981), p. 1218-1228.

[44] G. H. Golub, M. Heath et G. Wahba. "Generalized cross-validation as a method for choosing a good ridge parameter". In : *Technometrics* 21.2 (1979), p. 215-223.

[45] D. Haaland et E. Thomas. "Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information". In : *Analytical Chemistry* 60 (1988), p. 1193-1202.

[46] P. Hall. "Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density". In : *Ann. Statist.* 20.2 (1992), p. 675-694.

[47] J. Hanley et S. Shapiro. "Sexual Activity and the Lifespan of Male Fruitflies : A Dataset That Gets Attention". In : *J. of Statistics Education* 2.1 (1994). Données : `www-unix.oit.umass.edu/~statdata`.

[48] M. H. Hansen et B. Yu. "Model selection and the principle of minimum description length". In : *J. Amer. Statist. Assoc.* 96.454 (2001), p. 746-774.

[49] Heavenrich, Murrell et Hellman. "Light Duty Automotive Technology and Fuel Economy Trends Through 1991, U.S." In : *Environmental Protection Agency* EPA/AA/CTAB/91-02 (1991). Disponible par Internet sur DASL.

[50] M. Hofmann, C. Gatu et E. J. Kontoghiorghes. "Efficient algorithms for computing the best subset regression models for large-scale problems". In : *Comput. Statist. Data Anal.* 52.1 (2007), p. 16-29.

[51] M. Hout. "Status, Autonomy and Training in Occupational Mobility". In : *American J. of Sociology* 89.6 (1984), p. 1379-1409.

[52] P. J. Huber. *Robust statistics*. Wiley, 1981.

[53] H. Hundborg et al. "Familial Tendency to Fetal Loss". In : *Statistics in Medicine* 19 (2000), p. 2147-2168.

[54] R. Jenrich. "Asymptotic Properties of Non-linear Least Squares Estimators". In : 40 (1963), p. 633-643.

[55] P. Jones et G. McLachlan. "Fitting finite mxture models in a regression context". In : *Austral. J. Statist.* 34.2 (1992), p. 233-240.

[56] Y. Kim, H. Choi et H.-S. Oh. "Smoothly clipped absolute deviation on high dimensions". In : *J. Amer. Statist. Assoc.* 103.484 (2008), p. 1665-1673.

[57] C. Lavergne et C. Trottier. "Sur l'estimation dans les modèles linéaires généralisés à effets aléatoires". In : *Revue de Statistique Appliquée* 48.1 (2000), p. 49-67.

[58] R. Y. Liu. "Bootstrap procedures under some non-i.i.d. models". In : *Ann. Statist.* 16.4 (1988), p. 1696-1708.

[59] M. Mackisack. "What is the use of experiments conducted by statistics students ?" In : *J. of Stat. Educ.* 2.1 (1994). Données DASL : `www.statsci.org/data/oz/planes.html`.

[60] C. L. Mallows. "Augmented partial residuals". In : *Technometrics* 28.4 (1986), p. 313-319.

[61] E. Mammen. "Bootstrap, wild bootstrap, and asymptotic normality". In : *Probab. Theory Related Fields* 93.4 (1992), p. 439-455.

[62] R. Mazumder, J. H. Friedman et T. Hastie. "*SparseNet* : coordinate descent with nonconvex penalties". In : *J. Amer. Statist. Assoc.* 106.495 (2011), p. 1125-1138.

[63] M. Mougeot, D. Picard et K. Tribouley. "Learning out of leaders". In : *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 74.3 (2012), p. 475-513.

[64] R. Nuzzo. "Statistical errors". In : *Nature* 506 (2014), p. 150-152.

[65] L. Partridge et M. Farquhar. "Sexual Activity and the Lifespan of Male Fruitflies". In : *Nature* 294 (1981), p. 580-581.

[66] J. C. Pinheiro et D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[67] B. T. Polyak et A. B. Tsybakov. "A family of asymptotically optimal methods for selecting the order of a projection estimator for a regression". In : 37.3 (1992), p. 502-512.

[68] S. Press et S. Wilson. "Choosing Between Logistic Regression and Discriminant Analysis". In : *Journal of the American Statistical Association* 73.364 (1978), p. 699-705.

[69] O. Reiersøl. "Identifiability of a linear relation between variables which are subject to error". In : *Econometrica* 18 (1950), p. 375-389.

[70] J. Sacks et al. "Design and analysis of computer experiments". In : *Statist. Sci.* 4.4 (1989), p. 409-435.

[71] H. Scheffé. *The analysis of variance*. Wiley, 1959.

[72] G. Schwarz. "Estimating the dimension of a model". In : *Ann. Statist.* 6.2 (1978), p. 461-464.

[73] J. Shao. "Linear model selection by cross-validation". In : *J. Amer. Statist. Assoc.* 88.422 (1993), p. 486-494.

[74] J. Shao et X. Deng. "Estimation in high-dimensional linear models with deterministic design matrices". In : *Ann. Statist.* 40.2 (2012), p. 812-831.

[75] M. Solari. "The maximum likelihood solution of the problem of estimating a linear functional realtionship". In : *J. Roy. Statist. Soc. Ser. B* 31 (1969), p. 372-375.

[76] P. Tandeo et al. "Segmentation of Mesoscale Ocean Surface Dynamics Using Satellite SST and SSH Observations". In : *IEEE-GRS* (2013).

[77] P. F. Thall, K. E. Russell et R. M. Simon. "Variable selection in regression via repeated data splitting". In : *J. Comput. Graph. Statist.* 6 (1997), p. 1-34.

[78] S. Weisberg. *Applied linear regression*. Wiley, 1980.

[79] H. White. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". In : *Econometrica* 48.4 (1980), p. 817-838.

[80] World-Health-Organization. "Venous thromboembolic disease and combined oral contraceptives". In : *The Lancet* (1995), p. 1575-1582.

[81] T. Zhang. "Adaptive forward-backward greedy algorithm for learning sparse representations". In : *IEEE Trans. Inform. Theory* 57.7 (2011), p. 4689-4708.

Index