UNIVERSITE
PAUL
SABATIER
TOULOUSE III

PUBLICATIONS DU LABORATOIRE
DE
STATISTIQUE ET PROBABILITÉS

Pratique de la modélisation Statistique

PHILIPPE BESSE

Version janvier 2003 — mises à jour : `www.lsp.ups-tlse.fr/Besse`

Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Université Paul Sabatier — 31062 – Toulouse cedex 4.

Table des matières

Introduction

La *Statistique* a plusieurs objets : descriptif ou exploratoire, décisionnel (tests), modélisation selon que l'on cherche à représenter des structures des données, confirmer ou expliciter un modèle théorique ou encore prévoir. Ce cours s'intéresse au thème de la *modélisation* et plus particulièrement aux méthodes *linéaires* et à celles qui se ramènent au cas linéaire. Il se limite donc à l'exposé des méthodes dites *paramétriques* dans lesquelles interviennent des *combinaisons linéaires* des variables dites explicatives. Celles-ci visent donc à l'estimation d'un nombre généralement restreint de paramètres intervenant dans cette combinaison mais sans aborder les techniques spécifiques à l'étude des séries chronologiques. Les méthodes non-paramétriques élémentaires (loess, noyaux, splines) seront introduites dans le cas unidimensionnel.

Le *cadre général* de ce cours considère donc les observations d'une variable aléatoire Y dite *réponse, exogène, dépendante* qui doit être expliquée (modélisée) par les mesures effectuées sur p variables dites *explicatives, de contrôle, endogènes, dépendantes, régresseurs*. Ces variables peuvent être quantitatives ou qualitatives, ce critère déterminant le type de méthode ou de modèle à mettre en œuvre : régression linéaire, analyse de variance et covariance, régression logistique, modèle log-linéaire.

Compte tenu du temps limité et de la variété des outils mis en jeu nous avons fait le choix d'insister sur la *pratique* des méthodes considérées ainsi que sur la compréhension des sorties proposées par un logiciel (SAS/STAT) et de leurs limites plutôt que sur les fondements théoriques. Ce cours s'inspire largement d'une présentation "anglo-saxonne" de la Statistique, du particulier vers le général, dont des compléments sont à rechercher dans la bibliographie citée en référence. On montre donc comment utiliser les propriétés des modèles statistiques pour le traitement des données tandis que certains des aspects plus mathématiques (démonstrations) sont l'objet d'exercices. Néanmoins, le dernier chapitre introduit au cadre théorique général incluant toutes les méthodes considérées : le *modèle linéaire généralisé*.

En théorie, on peut distinguer deux approches : avec ou sans hypothèse probabiliste sur la distribution des observations ou des erreurs qui est, le plus souvent, l'hypothèse de *normalité*. En pratique, cette hypothèse n'est guère prouvable, les tests effectués sur les résidus estimés sont peu puissants. Cette hypothèse est néanmoins implicitement utilisée par les logiciels qui produisent systématiquement les résultats de tests. Plus rigoureusement, ces résultats sont justifiés par les propriétés des distributions asymptotiques des estimateurs, propriétés qui ne sont pas développées dans ce cours. En conséquence, du moment que les échantillons sont de taille "raisonnable", hypothèse on non de normalité, les distributions des estimateurs et donc les statistiques de test sont considérées comme valides.

En revanche, d'autres aspects des hypothèses, inhérentes aux méthodes développées et qui, en pratique, conditionnent fortement la qualité des estimations, doivent être évalués avec soin : *linéarité, colinéarité, homoscédasticité, points influents* ou atypiques (outliers). Les différents *diagnostics* ainsi que le problème du choix des variables explicatives, c'est-à-dire du *choix de modèle*, sont plus particulièrement décrits.

Dans la mesure du possible, nous avons respecté une certaine uniformisation des notations. Des caractères majuscules X, Y désignent des variables aléatoires, des caractères gras minuscules désignent des vecteurs : y_i est la $i$ème observation de Y rangée dans le vecteur $\mathbf{y}$, un chapeau désigne un prédicteur : $\hat{y}_i$, les caractères gras majuscules sont des matrices, un caractère grec (β) est un paramètre (qui est une variable aléatoire) dont l'estimation est désignée par la lettre latine correspondante (b).

Enfin, ce support de cours est et restera longtemps en chantier, les mises à jour successives ainsi que des sujets de travaux pratiques sont disponibles à partir de l'URL :

`www-sv.cict.fr/lsp/Besse`.

Chapitre 1

Régression linéaire simple

Ce chapitre élémentaire permet d'introduire simplement certains concepts clefs : modèle, estimations, tests, diagnostics, qui seront ensuite déclinés dans des cadres plus généraux. Il vient en complément d'un cours traditionnel de Statistique de niveau bac+3 sur l'estimation et les tests.

1 Modèle

On note Y la variable aléatoire réelle à expliquer et X la variable explicative (déterministe) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, $E(Y)$, est une fonction affine de X.

$$E(Y) = f(X) = \beta_0 + \beta_1 X.$$

Remarque : Nous supposerons pour simplifier que X est déterministe. Dans le cas contraire, X aléatoire, le modèle s'écrit alors conditionnellement aux observations de X : $E(Y|X = x) = \beta_0 + \beta_1 x$ et conduit aux mêmes estimations.

Pour une séquence d'observations aléatoires identiquement distribuées $\{(y_i, x_i)i = 1, \ldots, n\}$ ($n > 2$, et les x_i non tous égaux), le modèle s'écrit avec les observations :

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad i = 1, \ldots, n$$

ou sous la forme matricielle :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

où le vecteur $\mathbf{u}$ contient les erreurs.

Les *hypothèses* relatives à ce modèle sont les suivantes :

i. la distribution de l'erreur $\mathbf{u}$ est indépendante de X **ou** X est fixe,

ii. l'erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, \ldots, n \quad E(u_i) = 0, \quad \text{Var}(u_i) = \sigma_u^2.$$

iii. β_0 et β_1 sont constants, pas de rupture du modèle.

iv. Hypothèse complémentaire pour les inférences : $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$.

2 Estimation

L'estimation des paramètres $\beta_0, \beta_1, \sigma^2$ est obtenue en maximisant la vraisemblance, sous l'hypothèse que les erreurs sont gaussiennes, ou encore par minimisation de la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données $\{(x_i, y_i) i = 1 \dots, n\}$, le critère des moindres carrés s'écrit :

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

On pose :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$
$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad r = \frac{s_{xy}}{s_x s_y};$$

Les moindres carrés sont obtenus par :

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2}, \\ b_0 &= \bar{y} - b_1 \bar{x}. \end{aligned}$$

On montre que ce sont des estimateurs sans biais et de variance minimum parmi les estimateurs fonctions linéaires des y_i (resp. parmi tous les estimateurs dans le cas gaussien). À chaque valeur de X correspond la valeur *estimée* (ou prédite, ajustée) de Y :

$$\widehat{y_i} = b_0 + b_1 x_i,$$

les *résidus* calculés ou estimés sont :

$$e_i = y_i - \widehat{y_i}.$$

La variance σ_u^2 est estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

2.1 Inférence

Les estimateurs b_0 et b_1 sont des variables aléatoires réelles de matrice de covariance :

$$\sigma_u^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} & -\frac{\bar{x}}{(n-1)s_x^2} \\ -\frac{\bar{x}}{(n-1)s_x^2} & \frac{1}{(n-1)s_x^2} \end{bmatrix}$$

qui est estimée en remplaçant σ_u^2 par son estimation s^2. Sous l'hypothèse que les résidus sont gaussiens, on montre que

$$\frac{(n-2)s^2}{\sigma_u^2} \sim \chi^2_{(n-2)}$$

et donc que les statistiques

$$(b_0 - \beta_0) \Big/ s\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2} \qquad \text{et} \qquad (b_1 - \beta_1) \Big/ s\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}$$

suivent des lois de Student à $(n-2)$ degrés de liberté. Ceci permet de tester l'hypothèse de nullité d'un de ces paramètres ainsi que de construire les intervalles de confiance :

$$\begin{aligned} b_0 &\pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2}, \\ b_1 &\pm t_{\alpha/2;(n-2)} s \left(\frac{1}{(n-1)s_x^2} \right)^{1/2}. \end{aligned}$$

Attention : une inférence conjointe sur β_0 et β_1 ne peut être obtenue en considérant séparément les intervalles de confiance. La région de confiance est en effet une ellipse d'équation :

$$n(b_0-\beta_0)^2 + 2(b_0-\beta_0)(b_1-\beta_1)\sum_{i=1}^{n} x_i + (b_1-\beta_1)^2 \sum_{i=1}^{n} x_i^2 = 2s^2 \mathcal{F}_{\alpha;2,(n-2)}$$

qui est inclue dans le rectangle défini par les intervalles. Un grande part des valeurs du couple (β_0, β_1) est donc exclue de la région de confiance et ce d'autant plus que b_0 et b_1 sont corrélés.

3 Qualité d'ajustement, prédiction

Il est d'usage de décomposer les sommes de carrés des écarts à la moyenne sous la forme ci-dessous ; les notations sont celles de la plupart des logiciels :

$$\begin{array}{lll} \textit{Total sum of squares} & \text{SST} & = (n-1)s_y^2, \\ \textit{Regression sum of squares} & \text{SSR} & = (n-1)\frac{s_{xy}^2}{s_x^2}, \\ \textit{Error sum of squares} & \text{SSE} & = (n-2)s^2, \end{array}$$

et on vérifie : SST = SSR + SSE.

On appelle *coefficient de détermination* la quantité

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{n-2}{n-1}\frac{s^2}{s_y^2} = \frac{\text{SSR}}{\text{SST}}$$

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale.

Sous l'hypothèse : $\beta_1 = 0$, la statistique

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

suit une distribution de Fisher $\mathcal{F}_{1,(n-2)}$. Cette statistique est le carré de la statistique de Student correspondant à la même hypothèse.

Connaissant une valeur x_0, on définit deux *intervalles de confiance de prédiction* à partir de la valeur prédite $\widehat{y_0} = b_0 + b_1 x_0$. Le premier encadre $E(Y)$ sachant $X = x_0$; le deuxième, qui encadre $\widehat{y_0}$ est plus grand car il tient compte de la variance totale : $\sigma_u^2 + \text{Var}(\widehat{y_0})$:

$$\begin{aligned} \widehat{y_0} &\pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2}, \\ \widehat{y_0} &\pm t_{\alpha/2;(n-2)} s \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2}. \end{aligned}$$

Les logiciels proposent également une *bande de confiance* entre deux arcs d'hyperboles pour la droite de régression. À chaque point (b_0, b_1) de l'ellipse de confiance de (β_0, β_1) correspond une droite d'équation $\widehat{y} = b_0 + b_1 x$. Toutes ces droites sont comprises entre les bornes :

$$\widehat{y} \pm s\sqrt{\mathcal{F}_{1,(n-2)}} \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$

Ceci signifie que cette bande recouvre la “vraie” ligne avec une probabilité $1-\alpha$. Elle est plus grande que celle associée aux intervalles de confiance des $E(Y)$.

Attention : la prédiction par intervalle n’est justifiée que pour des observations appartenant à la population échantillonnée et à condition que les hypothèses : linéarité, erreurs i.i.d., (normalité), soient valides. Éviter les extrapolations.

4 Nuage de points, transformations

Toute tentative de modélisation nécessite une étude descriptive préalable afin de s’assurer, au moins graphiquement, de la validité des hypothèses considérées. Ceci passe

i. par une étude uni-variée de chaque distribution pour détecter des dissymétries ou encore des valeurs atypiques (outliers) : boîtes à moustaches, histogrammes, estimation non-paramétrique de la densité,
ii. puis par une représentation du nuage de points dans le repère (X, Y) et une régression non-paramétrique afin de déceler une éventuelle liaison non-linéaire entre les variables. *Attention*, même si elle est forte, une liaison non-linéaire, par exemple de type quadratique entre X et Y, peut conduire néanmoins a un coefficient de corrélation linéaire très faible.

Dans les deux cas, en cas de problèmes, le remède consiste souvent à rechercher des transformations des variables permettant de rendre les distributions symétriques, de “banaliser” les points atypiques et de rendre linéaire la relation. La qualité de l’estimation d’une distribution par un histogramme dépend beaucoup du découpage en classe. Malheureusement, plutôt que de fournir des classes d’effectifs égaux et donc de mieux répartir l’imprécision, les logiciels utilisent des classes d’amplitudes égales et tracent donc des histogrammes parfois peu représentatifs. Ces 20 dernières années, à la suite du développement des moyens de calcul, sont apparues des méthodes d’estimation dites *fonctionnelles* ou *non-paramétriques* qui proposent d’estimer la distribution d’une variable ou la relation entre deux variables par une fonction construite point par point (noyaux) ou dans une base de fonctions *splines*. Ces estimations sont simples à calculer (pour l’ordinateur) mais nécessitent le choix d’un paramètre dit de *lissage*. Les démonstrations du caractère optimal de ces estimations fonctionnelles, liée à l’optimalité du choix de la valeur du paramètre de lissage, font appel à des outils théoriques plus sophistiquées sortant du cadre de ce cours (Eubank 1988, Silverman 1986).

Nous résumons ci-dessous les techniques non-paramétriques, simples et efficaces dans ce genre de situation, trop rarement enseignées dans un cours de statistique descriptive, mais déjà présentes dans certains logiciels (SAS/INSIGHT).

4.1 Estimation de la densité

L’estimation de la densité par la méthode du noyau se met sous la forme générale :

$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x-x_i}{\lambda}\right)$$

où λ est le paramètre de lissage optimisée par une procédure automatique qui minimise une approximation de l’erreur quadratique moyenne intégrée (MISE : norme dans l’espace L^2) ; K est une fonction symétrique, positive, concave, appelée *noyau* dont la forme précise importe peu. C’est souvent la fonction densité de la loi gaussienne :

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$$

qui possède de bonnes propriétés de régularité. Le principe consiste simplement à associer à chaque observation un “élément de densité” de la forme du noyau K et à sommer tous ces éléments. Un histogramme est une version particulière d’estimation dans laquelle l”’élément de densité” est un “petit rectangle” dans la classe de l’observation.

4.2 Régression non-paramétrique

On considère un modèle de régression de la forme

$$y_i = f(x_i) + \varepsilon_i$$

où les erreurs sont centrées et la fonction f est supposée régulière : existence de dérivées jusqu'à un certain ordre. Dans ce contexte, de nombreux estimateurs de f ont été proposés. Ils conduisent souvent à des résultats assez voisins, le point le plus sensible étant le choix de λ.

Spline

Le *lissage spline* élémentaire consiste à rechercher, dans l'espace des fonctions continûment différentiables et avec une dérivée seconde de carré intégrable, le minimum d'un critère combinant ajustement des observations et régularité de la solution :

$$\widehat{f_\lambda} = \arg\min_f \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_{-\infty}^{+\infty} (f''(x))^2 dx.$$

On montre que l'on obtient une fonction polynômiale (de degré 3) par morceaux. La valeur optimale du paramètre de lissage est fixée par validation croisée généralisée (GCV).

Noyau

La régression non-paramétrique par la méthode du noyau consiste à calculer une moyenne pondérée autour de chaque observation. La pondération est fixée par une fonction K du même type que celle utilisée pour l'estimation de la densité.

$$\widehat{f_\lambda}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x-x_i}{\lambda}\right) x_i}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{\lambda}\right)}.$$

Loess

L'estimateur précédent est susceptible de biais même dans le cas simple de points alignés. Une adaptation propose de calculer, plutôt qu'une moyenne locale pondérée, une régression linéaire ou même quadratique locale. On parle alors de lisseur polynômial local.

transformations

Dans le cas où des problèmes (distribution, non-linéarité) ont été identifiés, l'étape suivante consiste à rechercher des transformations élémentaires (logarithme, puissance) des variables susceptibles de les résoudre. Ceci amène à étudier les modèles des exemples suivants :

$$\begin{aligned}
Y &= \beta_0 + \beta_1 \ln X \\
\ln Y &= \beta_0 + \beta_1 X \quad \text{ou} \quad Y = ab^X \quad \text{avec} \quad \beta_0 = \ln a \text{ et } \beta_1 = \ln b \\
\ln Y &= \beta_0 + \beta_1 \ln X \quad \text{ou} \quad Y = aX^{\beta_1} \quad \text{avec} \quad \beta_0 = \ln a \\
Y &= \beta_0 + \beta_1 (1/X) \\
Y &= \beta_0 + \beta_1 X^{1/2} \\
Y &= \beta_0 + \beta_1 X^2 \quad \text{ou, plus généralement,} \\
Y &= \beta_0 + \beta_1 X^\alpha \\
&\ldots
\end{aligned}$$

5 Influence

Le critère des moindres carrés, comme la vraisemblance appliquée à une distribution gaussienne douteuse, est très sensible à des observations atypiques, hors "norme" (outliers) c'est-à-dire qui présentent des valeurs trop singulières. L'étude descriptive initiale permet sans doute déjà d'en repérer mais c'est insuffisant. Un diagnostic doit être établi dans le cadre spécifique du modèle recherché afin d'identifier les observations *influentes* c'est-à-dire celles dont une faible variation du couple (x_i, y_i) induisent une modification importante des caractéristiques du modèle.

Ces observations repérées, il n'y a pas de remède universel : supprimer un valeur aberrante, corriger une erreur de mesure, construire une estimation robuste (en norme L_1), ne rien faire..., cela dépend du contexte et doit être négocié avec le commanditaire de l'étude.

5.1 Effet levier

Une première indication est donnée par l'éloignement de x_i par rapport à la moyenne $\bar{x}$. En effet, écrivons les prédicteurs $\widehat{y_i}$ comme combinaisons linéaires des observations (cf. exo 3) :

$$\widehat{y_i} = b_0 + b_1 x_i = \sum_{j=1}^{n} h_{ij} y_j \quad \text{avec} \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \; ;$$

en notant $\mathbf{H}$ la matrice (hat matrix) des h_{ij} ceci s'exprime encore matriciellement :

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation de $\widehat{y_i}$.

5.2 Résidus

Différents types de résidus sont définis afin d'affiner leurs propriétés.

Résidus : $e_i = y_i - \widehat{y_i}$

Résidus$_i$: $e_{(i)i} = y_i - \widehat{y_{(i)i}} = \frac{e_i}{1-h_{ii}}$
où $\widehat{y_{(i)i}}$ est la prévision de y_i calculée sans la $i$ème observation (x_i, y_i). On note

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)i}^2 \quad \text{(predicted residual sum of squares)}$$

la somme des carrés de ces résidus.

Résidus standardisés : Même si l'hypothèse d'homoscédasticité est vérifiée, ceux-ci n'ont pas la même variance : $E(e_i) = 0$ et $\text{Var}(e_i) = \sigma_u^2(1 - h_{ii})$. Il est donc d'usage d'en calculer des versions *standardisées* afin de les rendre comparables :

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}.$$

Résidus studentisés : La standardisation ("interne") dépend de e_i dans le calcul de s estimation de $\text{Var}(e_i)$. Une estimation non biaisée de cette variance est basée sur

$$s_{(i)}^2 = \left[(n-2)s^2 - \frac{e_i^2}{1 - h_{ii}}\right] / (n-3)$$

qui ne tient pas compte de la $i$ème observation. On définit alors les résidus *studentisés* par :

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}.$$

Sous hypothèse de normalité, on montre que ces résidus suivent une loi de Student à $(n-3)$ degrés de liberté.

Il est ainsi possible de construire un test afin tester la présence d'*une* observation atypique ou de plusieurs en utilisant l'inégalité de Bonferroni. Plus concrètement, en pratique, les résidus studentisés sont comparés aux bornes ± 2.

5.3 Diagnostics

Les deux critères précédents contribuent à déceler des observations potentièlement influentes par leur éloignement à $\bar{x}$ ou la taille des résidus. Ces informations sont synthétisées dans des critères évaluant directement l'influence d'une observation sur certains paramètres : les prédictions $\widehat{y_i}$, les paramètres b_0, b_1, le déterminant de la matrice de covariance des estimateurs. Tous ces indicateurs proposent de comparer un paramètre estimé sans la $i$ème observation et ce même paramètre estimé avec toutes les observations.

Le plus couramment utilisé est la distance de Cook :

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y_{(i)j}} - \widehat{y_j})^2}{2s^2} = \frac{h_{ii}}{2(1-h_{ii})} r_i^2 \qquad \text{pour} \quad i = 1, \ldots, n$$

qui mesure donc l'influence d'une observation sur l'ensemble des prévisions en prenant en compte effet levier et importance des résidus.

La stratégie de détection consiste le plus souvent à repérer les points atypiques en comparant les distances de Cook avec la valeur 1 puis à expliquer cette influence en considérant, pour ces observations, leur résidu ainsi que leur effet levier.

6 Graphe des résidus

Le nuage des points (x_i, y_i) assorti d'un lissage permet de détecter une éventuelle relation non-linéaire entre les variables. D'autres hypothèses doivent être validées :

- l'homoscédasticité par un graphique des résidus studentisés ou non : (x_i, t_i) afin de repérer des formes suspectes de ce nuage qui devrait se répartir uniformément de part et d'autre de l'axe des abscisses,
- éventuellement la normalité des résidus en étudiant leur distribution,
- l'autocorrélation des résidus dans le cas, par exemple, où la variable explicative est le temps.

Une transformation des variables ou une modélisation spécifique à une série chronologique (SARIMA) permet, dans les situations favorables, de résoudre les difficultés évoquées.

7 Exemple

Pour 47 immeubles d'appartements locatifs d'une grande ville américaine, les données (Jobson, 1991) fournissent le "revenu net" en fonction du "nombre d'appartements". Les tableaux ci-dessous sont des extraits des résultats fournis par la procédure `reg` du module SAS/STAT. Cette procédure génère beaucoup d'autres résultats comme les matrices $\mathbf{X}'\mathbf{X}$ (crossproducts), $\mathbf{X}'\mathbf{D}\mathbf{X}$ (model crossproducts) et son inverse, matrices des variances et corrélations des estimateurs.

```
proc reg data=sasuser.suitinco all;
model revenu=nbappart /dw Influence cli clm;
output out=hubout h=lev p=pred r=res student=resstu ;
run;
```

```
Descriptive Statistics
Variables    Sum        Mean      Uncorrected SS    Variance    Std Deviation
INTERCEP      47              1             47               0              0
NBAPPART    1942   41.319148936         157970    1689.7437558   41.106492866
REVENU   4336086   92257.148936   947699637616     11905754472   109113.49354
Correlation : 0.8856

Analysis of Variance
                          Sum of         Mean
Source          DF      Squares       Square           F Value      Prob>F
                (1)
Model            1 429511948724 (2) 429511948724 (5)  163.585 (7)  0.0001 (8)
Error           45 118152756990 (3)  2625616822  (6)
C Total         46 547664705714 (4)

   Root MSE   51240.77304 (9)    R-square       0.7843   (12)
   Dep Mean   92257.14894 (10)   Adj R-sq       0.7795
   C.V.          55.54125 (11)
```

(1)	degrés de liberté de la loi de Fisher du test global ($H_0 : \beta_1 = 0$)
(2)	SSR
(3)	SSE ou déviance
(4)	SST=SSE+SSR
(5)	SSR/DF
(6)	s^2 =MSE=SSE/DF est l'estimation de σ_u^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10) sans échelle ni dimension
(12)	Coefficient de détermination R^2 ou carré du coefficient de corrélation.

Parameter Estimates

Variable	DF	Parameter Estimate (1)	Standard Error (2)	T for H0: Parameter=0 (3)	Prob > |T| (4)
INTERCEP	1	-4872.015285	10655.276212	-0.457	0.6497
NBAPPART	1	2350.705828	183.79188506	12.790	0.0001

(1)	estimations des paramètres (b_j)
(2)	écarts-types de ces estimations (s_{b_j})
(3)	statistique T du test de Student de $H_0 : b_j = 0$ $((b_j - 0)/s_{b_j})$
(4)	$P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$

Connaissant les fractiles de la loi de Student : $t_{0,975;45} = 2,015$, on construit facilement des intervalles de confiance des estimateurs, ici au niveau 5% : $[b_j - t_{0,975;n-2}s_{b_j}; b_j + t_{0,975;n-2}s_{b_j}]$.

Obs	Dep Var REVENU (1)	Predict Value (2)	Std Err Predict (3)	Lower95 Mean (4)	Upper95 Mean (5)	Lower95 Predict (6)	Upper95 Predict (7)	Residual (8)	Std Err Resid. (9)	Student Resid. (10)
1	119202	131469	8078.5	115198	147740	26989.9	235948	-12266.9	50599.9	-0.242
...										
23	345608	239601	13732.	211943	267260	132755	346448	106007	49366.3	2.147
24	350633	324227	19616.	284717	363736	213718	434735	26406.2	47337.2	0.558
25	226375	98559.	7490.4	83472.	113646	-5742.0	202860	127816	50690.3	2.522
26	247203	178483	10065.	158210	198756	73306.5	283660	68720.0	50242.4	1.368
27	28519.	157327	9041.4	139116	175537	52528.2	262125	-128808	50436.7	-2.55
28	154278	347734	21327.	304779	390689	235947	459521	-193456	46591.4	-4.15
29	157332	140872	8385.2	123983	157760	36294.8	245449	16460.3	50550.0	0.326
30	171305	197289	11104.	174924	219653	91689.0	302888	-25983.7	50023.1	-0.52
...										

Obs	-2-1-0 1 2 (11)	Cook's D (12)	Rstudent (13)	Hat Diag H (14)	Cov Ratio (15)	Dffits (15)	INTERCEP Dfbetas (15)	NBAPPART Dfbetas (15)
1	| |	0.001	-0.2399	0.0249	1.0698	-0.0383	-0.0145	-0.0145
...								
23	| |**** |	0.178	2.2413	0.0718	0.9078	0.6235	-0.1347	0.5230
24	| |* |	0.027	0.5535	0.1466	1.2087	0.2294	-0.0898	0.2121
25	| |***** |	0.069	2.6906	0.0214	0.7881	0.3976	0.2597	0.0262
26	| |** |	0.038	1.3815	0.0386	0.9994	0.2768	0.0120	0.1854
27	| *****| |	0.105	-2.7310	0.0311	0.7893	-0.4896	-0.0876	-0.2755
28	|******| |	1.806	-5.2275	0.1732	0.4814	-2.3929	1.0090	-2.2411
29	| | |	0.001	0.3224	0.0268	1.0697	0.0535	0.0162	0.0242
30	| *| |	0.007	-0.5152	0.0470	1.0844	-0.1144	0.0063	-0.0846
...								

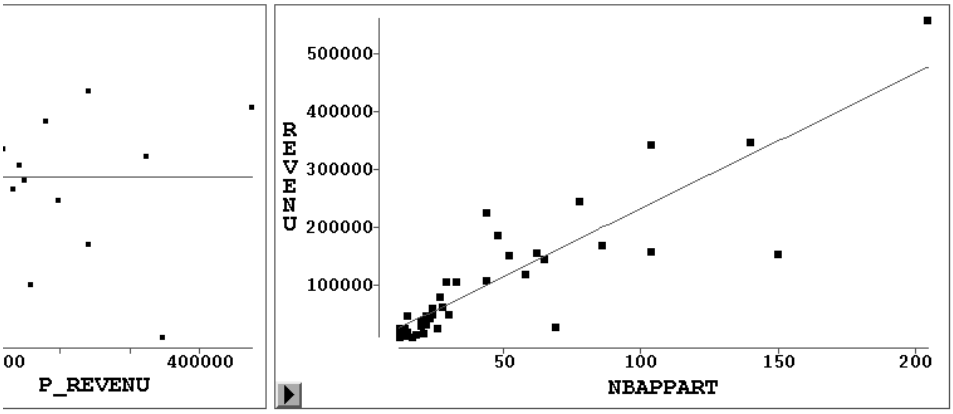


FIG. 1.1 – Graphe des résidus et nuage de points de la régression du revenu en fonction du nombre d'appartements.

(1)	variable à expliquer y_i
(2)	valeur ajustée $\widehat{y_i}$
(3)	écart-type de cette estimation$s_{\widehat{y_i}}$
(4)et (5)	Intervalle de confiance pour l'estimation de $E(y_i)$
(6) et (7)	Intervalle de confiance pour l'estimation de y_i
(8)	résidus calculés $e_i = y_i - \widehat{y_i}$
(9)	écarts-types de ces estimations
(10)	résidus standardisés (ou studentisés internes) r_i
(11)	repérage graphique des résidus standardisés : $* = 0.5$.
(12)	Distance de Cook
(13)	résidus studentisés (externes) t_i
(14)	Termes diagonaux de la matrice chapeau $\mathbf{H}$
(15)	autres indicateurs d'influence

Les observations 28 et 16 seraient à inspecter avec attention. Certaines, dont la 28, présentent une valeur observée hors de l'intervalle de prédiction.

Le graphique des résidus sont présentés dans la figure 1.1. Il montre clairement que l'hypothèse d'homoscédasticité n'est pas satisfaite. Une autre modélisation faisant intervenir une transformation des variables serait nécessaire. Ainsi la modélisation du logarithme du revenu en fonction du logarithme du nombre d'appartements représentée par la figure 1.2 est nettement plus satisfaisante. Une étude descriptive préalable des distributions aurait permis de conduire à ce choix.

8 Exercices

Exo 1

Optimiser les moindres carrés de la section 2 pour retrouver les estimations des paramètres du modèle de régression simple.

Exo 2

Avec les notations précédentes relatives à la régression linéaire simple de Y sur X à partir des observations (x_i, y_i), montrer que

i. le coefficient de corrélation r^2 =SSR/SST,

ii. SST = SSE + SSR,

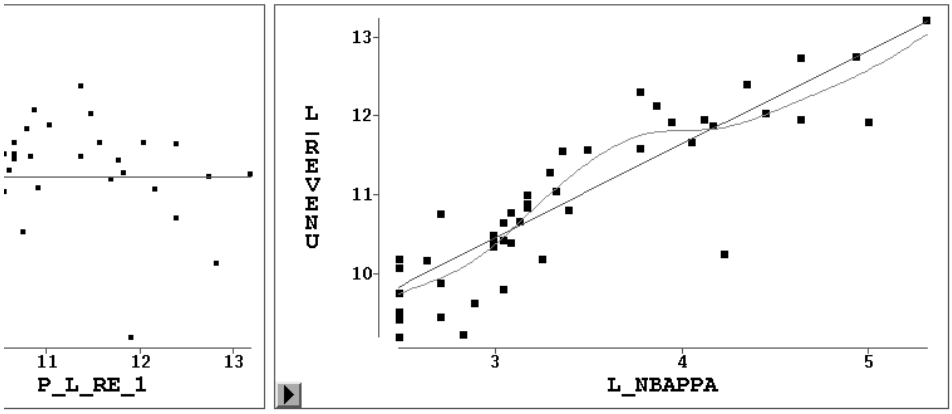


FIG. 1.2 – Graphe des résidus et nuage de points de la régression (linéaire et non paramétrique) du logarithme du revenu en fonction du logarithme du nombre d'appartements.

iii. $s^2 = \frac{n-1}{n-2} s_y^2 (1 - r^2)$.

Exo 3

on considère la régression linéaire simple de Y sur X à partir des observations (x_i, y_i).

i. Montrer que $\widehat{y_i}$ se met sous la forme

$$\widehat{y_i} = \sum_{j=1}^{n} h_{ij} y_j \qquad \text{avec} \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

ii. Posons $\mathbf{X} = [\mathbf{1}\mathbf{x}]$ la matrice $(n \times 2)$ contenant une colonne de 1 et le vecteur colonne des x_i. Calculer $\mathbf{X}'\mathbf{X}$, $(\mathbf{X}'\mathbf{X})^{-1}$ et la matrice $\mathbf{H}$ de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace engendré par les colonnes de $\mathbf{X}$.

iii. Calculer le terme général de cette matrice $\mathbf{H}$, en déduire que le vecteur $\widehat{\mathbf{y}}$ est obtenu par projection par $\mathbf{H}$ de $\mathbf{y}$.

iv. Calculer la covariance des $\widehat{y_i}$.

Exo 4

Dans le cadre de la régression simple, on considère les quantités $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}$ ainsi que celles $\bar{x}_{(i)}, \bar{y}_{(i)}, s_{x(i)}^2, s_{y(i)}^2, s_{xy(i)}$, calculées sans la $i$ème observation.

i. Montrer que

$$\begin{aligned} s_x^2 &= \frac{n-2}{n-1} s_{x(i)}^2 + \frac{1}{n} (\bar{x}_{(i)} - x_i)^2 \\ s_{xy} &= \frac{n-2}{n-1} s_{xy(i)} + \frac{1}{n} (\bar{x}_{(i)} - x_i)(\bar{y}_{(i)} - y_i). \end{aligned}$$

ii. En déduire les expressions de $s_{xy(i)}$ et $s_{x(i)}^2$ en fonction de $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}$.

Chapitre 2

Régression linéaire multiple

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

1 Modèle

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

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

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

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

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

avec les hypothèses suivantes :

i. Les u_i sont des termes d'erreur, d'une variable U, non observés, indépendants et identiquement distribués ; $E(u_i) = 0, Var(U) = \sigma_u^2 \mathbf{I}$.

ii. Les termes x^j sont supposés déterministes (facteurs contrôlés) **ou bien** l'erreur U est indépendante de la distribution conjointe de $X^1, \ldots, X^p$. On écrit dans ce dernier cas que :

$$E(Y|X^1, \ldots, X^p) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p \text{ et } Var(Y|X^1, \ldots, X^p) = \sigma_u^2.$$

iii. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.

iv. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur U ($\mathcal{N}(0, \sigma_u^2 \mathbf{I})$). Les u_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma_u^2)$.

Les données sont rangées dans une matrice $\mathbf{X}(n \times (p+1))$ de terme général x_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($x_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général y_i. En notant les vecteurs $\mathbf{u} = [u_1 \cdots u_p]'$ et $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$, le modèle s'écrit matriciellement :

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

2 Estimation

Conditionnellement à la connaissance des valeurs des X^j, les paramètres inconnus du modèle : le vecteur β et σ_u^2 (paramètre de nuisance), sont estimés par minimisation du critère des moindres carrés (M.C.)

ou encore, en supposant (iv), par maximisation de la vraisemblance (M.V.). Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

2.1 Estimation par M.C.

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

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

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

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

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

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

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

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

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

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

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

On note

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

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

2.2 Propriétés

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

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

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

celle des prédicteurs est

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

et celle des estimateurs des résidus est

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

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

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

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

2.3 Sommes des carrés

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

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

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

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

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

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

On vérifie alors : SST = SSR + SSE.

2.4 Coefficient de détermination

On appelle *coefficient de détermination* le rapport

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

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

Attention, dans le cas extrême où $n = (p + 1)$, c'est-à-dire si le nombre de variables explicatives est grand comparativement au nombre d'observations, $R^2 = 1$. Ou encore, il est géométriquement facile de voir que l'ajout de variables explicatives ne peut que faire croître le coefficient de détermination.

La quantité R est appelée *coefficient de corrélation multiple* entre Y et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{y}$ et sa prédiction (ou projection) $\widehat{\mathbf{y}}$.

3 Inférences dans le cas gaussien

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

3.1 Inférence sur les coefficients

Pour chaque coefficient β_j on montre que la statistique

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

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

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

Attention, cette statistique concerne un coefficient et ne permet pas d'inférer conjointement (cf. §3.4) sur d'autres coefficients car ils sont corrélés entre eux ; de plus elle dépend des absences ou présences des autres variables X^k dans le modèle. Par exemple, dans le cas particulier de deux variables X^1 et X^2 très corrélées, chaque variable, en l'absence de l'autre, peut apparaître avec un coefficient significativement différent de 0 ; mais, si les deux sont présentes dans le modèle, elles peuvent chacune apparaître avec des coefficients insignifiants.

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

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

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

3.2 Inférence sur le modèle

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

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

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

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

3.3 Inférence sur un modèle réduit

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

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

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

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

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

3.4 Ellipsoïde de confiance

Les estimateurs des coefficients β_j étant corrélés, la recherche d'une région de confiance de niveau $100(1-\alpha)\%$ pour tous les coefficients conduit à considérer l'ellipsoïde décrit par

$$(\mathbf{b} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}) \leq (p+1)s^2 F_{\alpha;p+1,(n-p-1)}.$$

Plus généralement, un ellipsoïde de confiance conjoint à q combinaisons linéaires $\mathbf{T}\boldsymbol{\beta}$ est donné par

$$(\mathbf{T}\mathbf{b} - \mathbf{T}\boldsymbol{\beta})'[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}(\mathbf{T}\mathbf{b} - \mathbf{T}\boldsymbol{\beta}) \leq qs^2 F_{\alpha;q,(n-p-1)}$$

où $\mathbf{T}(q \times (p+1))$ est une matrice de rang q de constantes fixées.

En application, étant donnés une matrice $\mathbf{T}$ et un vecteur $\mathbf{a}$, un test de l'hypothèse H_0 : $\mathbf{T}\beta = \mathbf{a}$ est obtenu en considérant la statistique

$$(\mathbf{T}\mathbf{b} - \mathbf{a})'[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}(\mathbf{T}\mathbf{b} - \mathbf{a})/qs^2$$

qui suit sous H_0 une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

3.5 Prévision

Connaissant les valeurs des variables X^j pour une nouvelle observation : $\mathbf{x}_0' = [x_0^1, x_0^2, \ldots, x_0^p]$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y}_0$ de Y ou $E(Y)$ est donnée par :

$$\widehat{y}_0 = b_0 + b_1 x_0^1 + \cdots + b_p x_0^p.$$

Les intervalles de confiance des prévisions de Y et $E(Y)$, pour *une* valeur $\mathbf{x}_0 \in \mathbb{R}^p$ et en posant $\mathbf{v}_0 = (1|bmx_0')' \in \mathbb{R}^{p+1}$, sont respectivement

$$\begin{aligned} \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)} s(1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}, \\ \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)} s(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}. \end{aligned}$$

Enfin, un intervalle de confiance de niveau $100(1-\alpha)\%$ recouvrant globalement la surface de régression est donné par

$$\widehat{y}_0 \pm [(p+1)F_{\alpha;(p+1),(n-p-1)}]^{1/2} s(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}.$$

Il peut être utilisé pour définir un intervalle conjoint à plusieurs prédictions.

4 Sélection de variables, choix de modèle

De façon un peu schématique, on peut associer la pratique de la modélisation statistique à trois objectifs qui peuvent éventuellement être poursuivis en complémentarité.

Descriptif : Il vise à rechercher de façon exploratoire les liaisons entre Y et d'autres variables, potentiellement explicatives, X^j qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. À cette stratégie, à laquelle peuvent contribuer des Analyses en Composantes Principales, correspond des algorithmes de recherche (pas à pas) moins performants mais économiques en temps de calcul si p est grand.

Attention, si n est petit, et la recherche suffisamment longue avec beaucoup de variables explicatives, il sera toujours possible de trouver un "bon" modèle expliquant y ; c'est l'effet *data mining* dans les modèles économétriques.

Explicatif : Le deuxième objectif est sous-tendu par une connaissance *a priori* du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels précédents permettent de construire *le* bon test conduisant à la prise de décision recherchée. Utilisées hors de ce contexte, les statistiques de test n'ont plus alors qu'une valeur indicative au même titre que d'autres critères plus empiriques.

Prédictif : Dans le troisième cas, l'accent est mis sur la qualité des estimateurs et des prédicteurs qui doivent, par exemple, minimiser une erreur quadratique moyenne. Ceci conduit à rechercher des modèles *parcimonieux* c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Le "meilleur" modèle ainsi obtenu peut donner des estimateurs légèrement biaisés au profit d'un compromis pour une variance plus faible. Un bon modèle n'est donc plus celui qui explique le mieux les données au sens d'une déviance (SSE) minimale (ou d'un R^2 max) au prix d'un nombre important de variables pouvant introduire des colinéarités. Le bon modèle est celui qui conduit aux prédictions les plus fiables.

4.1 Critères

De nombreux critères de choix de modèle sont présentés dans la littérature sur la régression linéaire multiple. Citons le critère d'information d'Akaïke (`AIC`), celui bayésien de Sawa (`BIC`), l'erreur quadratique moyenne de prédiction (cas gaussien)... Ils sont équivalents lorsque le nombre de variables à sélectionner, ou niveau du modèle, est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de niveaux différents. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression pénalisée de la fonction de vraisemblance afin de favoriser des modèles parcimonieux. En pratique, les plus utilisés ou ceux généralement fournis par les logiciels sont les suivants.

Statistique du F de Fisher

Ce critère, justifié dans le cas explicatif est aussi utilisé à titre indicatif pour comparer des séquences de modèles emboîtés. La statistique partielle de Fisher est

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

dans laquelle l'indice q désigne les expressions concernant le modèle réduit avec $(p-q)$ variables explicatives. On considère alors que si l'accroissement $(R^2 - R_q^2)$ est suffisamment grand :

$$R^2 - R_q^2 > \frac{q(1-R^2)}{(n-p-1)} F_{\alpha;q,(n-p-1)},$$

l'ajout des q variables au modèle est justifié.

R^2 et R^2 ajusté

Le coefficient de détermination $R^2 = 1-$SSE/SST, directement lié à la déviance (SSE) est aussi un indice de qualité mais qui a la propriété d'être monotone croissant en fonction du nombre de variables. Il ne peut donc servir qu'à comparer deux modèles de même niveau c'est-à-dire avec le même nombre de variables.

En revanche, le R^2 *ajusté* :

$$R'^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}.$$

dans lequel le rapport SSE/SST est remplacé par un rapport des estimations sans biais des quantités σ_u^2 et σ_y^2 introduit une pénalisation liée au nombre de paramètres à estimer.

Ce coefficient s'exprime encore par

$$1 - \frac{(n-1)\text{MSE}}{\text{SST}}$$

ainsi dans la comparaison de deux modèles partageant la même SST, on observe que $R'^2 > R'^2_j$ si et seulement si MSE$<$MSE$_j$; MSE et MSE$_j$ désignant respectivement l'erreur quadratique moyenne du modèle complet et celle d'un modèle à j variables explicatives. Maximiser le R^2 ajusté revient donc à minimiser l'erreur quadratique moyenne.

C_p de Mallows

Une erreur quadratique moyenne s'écrit comme la somme d'une variance et du carré d'un biais. L'erreur quadratique moyenne de prédiction sécrit ainsi :

$$\text{MSE}(\widehat{y}_i) = \text{Var}(\widehat{y}_i) + [\text{Biais}(\widehat{y}_i)]^2$$

puis après sommation et réduction :

$$\frac{1}{\sigma_u^2}\sum_{i=1}^{n}\text{MSE}(\widehat{y}_i) = \frac{1}{\sigma_u^2}\sum_{i=1}^{n}\text{Var}(\widehat{y}_i) + \frac{1}{\sigma_u^2}\sum_{i=1}^{n}[\text{Biais}(\widehat{y}_i)]^2.$$

En supposant que les estimations du modèle complet sont sans biais et en utilisant des estimateurs de $Var(\widehat{y}_i)$ et σ_u^2, l'expression de l'erreur quadratique moyenne totale standardisée (ou réduite) pour un modèle à q variables explicatives s'écrit :

$$C_p = (n-q-1)\frac{\text{MSE}_q}{\text{MSE}} - [n-2(q+1)]$$

et définit la valeur du C_p de Mallow pour les q variables considérées. Il est alors d'usage de rechercher un modèle qui minimise le C_p tout en fournissant une valeur inférieure et proche de $(q+1)$. Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise.

PRESS de Allen

On désigne par $\widehat{y}_{(i)}$ la prédiction de y_i calculée sans tenir compte de la $i$ème observation $(y_i, x_i^1, \ldots, x_i^p)$, la somme des erreurs quadratiques de prédiction (PRESS) est définie par

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2$$

et permet de comparer les capacités prédictives de deux modèles.

4.2 Algorithmes de sélection

Lorsque p est grand, il n'est pas raisonnable de penser explorer les 2^p modèles possibles afin de sélectionner le "meilleur" au sens de l'un des critères ci-dessus. Différentes stratégies sont donc proposées qui doivent être choisies en fonction de l'objectif recherché et des moyens de calcul disponibles ! Trois types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire c'est-à-dire par nombre croissant de modèles considérés parmi les 2^p et donc par capacité croissante d'optimalité. On donne pour chaque algorithme l'option `selection` à utiliser dans la procédure `REG` de SAS.

Pas à pas

Sélection (`forward`) À chaque pas, une variable est ajoutée au modèle. C'est celle dont la valeur p ("prob value")associée à la statistique partielle du test de Fisher qui compare les deux modèles est minimum. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque p reste plus grande qu'une valeur seuil fixée par défaut à $0,50$.

Élimination (`backward`) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable associée à la plus grande valeur p est éliminée du modèle. La procédure s'arrête lorsque les variables restant dans le modèle ont des valeurs p plus petites qu'un seuil fixé par défaut à $0,10$.

Mixte (`stepwise`) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Par échange

Maximisation de R^2 (`maxr`) Cet algorithme tente de trouver le meilleur modèle pour chaque niveau c'est-à-dire pour chaque nombre de variables explicatives. À chaque niveau il commence par sélectionner une variable complémentaire qui rend l'accroissement de R^2 maximum. Puis il regarde tous les échanges possibles entre une variable présente dans le modèle et une extérieure et exécute celui qui fournit l'accroissement maximum ; ceci est itéré tant que le R^2 croit.

Minimisation de R^2 (`minr`) Il s'agit du même algorithme que le précédent sauf que la procédure d'échange fait appel au couple de variables associé au plus petit accroissement du R^2. L'objectif est ainsi d'explorer plus de modèles que dans le cas précédent et donc, éventuellement, de tomber sur un meilleur optimum.

Remarque Pour tous ces algorithmes de sélection ou d'échange, il est important de compléter les comparaisons des différentes solutions retenues à l'aide de critères globaux (C_p ou PRESS).

Global

L'algorithme de Furnival et Wilson est utilisé pour comparer tous les modèles possibles en cherchant à optimiser l'un des critères : R^2, R^2 ajusté, ou C_p de Mallow (`rsquare, adjrsq, cp`) choisi par l'utilisateur. Par souci d'économie, cet algorithme évite de considérer des modèles de certaines sous-branches de l'arborescence dont on peut savoir a priori qu'ils ne sont pas compétitifs. En général les logiciels exécutant cet algorithme affichent le (`best=1`) ou les meilleurs modèles de chaque niveau.

5 Multi-colinéarité

L'estimation des paramètres ainsi que celle de leur écart-type (standard error) nécessite le calcul explicite de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$. Dans le cas dit *mal conditionné* où le déterminant de la matrice $\mathbf{X}'\mathbf{X}$ n'est que légèrement différent de 0, les résultats conduiront à des estimateurs de variances importantes et même, éventuellement, à des problèmes de précision numérique. Il s'agit donc de diagnostiquer ces situations critiques puis d'y remédier. Dans les cas descriptif ou prédictif on supprime des variables à l'aide des procédures de choix de modèle mais, pour un objectif explicatif nécessitant toutes les variables, d'autres solutions doivent être envisagées : algorithme de résolution des équations normales par transformations orthogonales (procédure `orthoreg` de SAS) sans calcul explicite de l'inverse pour limiter les problèmes numériques, régression biaisée (ridge), régression sur composantes principales.

5.1 Diagnostics

Notons $\widetilde{\mathbf{X}}$ la matrice des données observées, c'est-à-dire $\mathbf{X}$ privée de la première colonne $\mathbf{1}$ et dont on a retranché à chaque ligne le vecteur moyen $\overline{\mathbf{x}} = 1/n \sum_{i=1}^{n} \mathbf{x}_i$, $\mathbf{S}$ la matrice diagonale contenant les écarts-types empiriques des variables X^j et enfin $\mathbf{R}$ la matrice des corrélations :

$$\mathbf{R} = \frac{1}{(n-1)} \mathbf{S}^{-1} \widetilde{\mathbf{X}}' \widetilde{\mathbf{X}} \mathbf{S}^{-1}.$$

Facteur d'inflation de la variance (VIF)

Avec ces notations, la matrice de covariance des estimateurs des coefficients $(\beta_1, \ldots, \beta_p)$ s'écrit :

$$\frac{\sigma_u^2}{n-1} (\widetilde{\mathbf{X}}' \widetilde{\mathbf{X}})^{-1} = \frac{\sigma_u^2}{n-1} \mathbf{S} \mathbf{R}^{-1} \mathbf{S}.$$

On montre alors que chaque élément diagonal s'exprime comme

$$V_j = \frac{1}{1 - R_j^2}$$

où R_j^2 désigne le coefficient de détermination de la régression de la variable X^j sur les autres variables ; R_j est alors un coefficient de corrélation multiple, c'est le cosinus de l'angle dans $\mathbb{R}^n$ entre X^j et le sous-espace vectoriel engendré par les variables $\{X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p\}$. Plus X^j est "linéairement" proche de ces variables et plus R_j est proche de 1 et donc plus la variance de l'estimateur de β_j est élevée ; V_j est appelé *facteur d'inflation de la variance* (VIF). Évidemment, cette variance est minimum lorsque X^j est orthogonal au sous-espace engendré par les autres variables.

Le simple examen de la matrice $\mathbf{R}$ permet de relever des corrélations dangereuses de variables deux à deux mais est insuffisant pour détecter des corrélations plus complexes ou multi-colinéarités. C'est donc l'inverse de cette matrice qu'il faut considérer en calculant les V_j ou encore les valeurs $(1 - R_j^2)$ qui sont appelées *tolérances*.

Conditionnement

On note $\lambda_1, \ldots, \lambda_p$ les valeurs propres de la matrice $\mathbf{R}$ rangées par ordre décroissant. Le déterminant de $\mathbf{R}$ est égal au produit des valeurs propres. Ainsi, des problèmes numériques, ou de variances excessives apparaissent dès que les dernières valeurs propres sont relativement trop petites.

On appelle *indice de conditionnement* le rapport

$$\kappa = \lambda_1 / \lambda_p$$

de la plus grande sur la plus petite valeur propre.

En pratique, si $\kappa < 100$ on considère qu'il n'y a pas de problème. Celui-ci devient sévère pour $\kappa > 1000$. Cet indice de conditionnement donne un aperçu global des problèmes de colinéarité tandis que les VIF, les tolérances ou encore l'étude des vecteurs propres associés au plus petites valeurs propres permettent d'identifier les variables les plus problématiques.

Remarque : Lorsque le modèle est calculé avec un terme constant, la colonne $\mathbf{1}$ joue le rôle d'une variable et peut considérablement augmenter les problèmes de multi-colinéarité. La matrice $\mathbf{R}$ est alors remplacée par la matrice $\mathbf{T} = \text{diag}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{X}\text{diag}(\mathbf{X}'\mathbf{X})^{-1/2}$ dans les discussions précédentes.

5.2 Régression "ridge"

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur légèrement biaisé des paramètres. L'estimateur "ridge" introduisant une *régularisation* est donné par

$$\mathbf{b}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},$$

qui a pour effet de décaler de la valeur k toutes les valeurs propres de la matrice à inverser et, plus particulièrement, les plus petites qui reflètent la colinéarité. On montre que l'erreur quadratique moyenne sur l'estimation des paramètres se met sous la forme :

$$\text{MSE}(\mathbf{b}_R) = \sigma_u^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\boldsymbol{\beta}.$$

La difficulté est alors de trouver une valeur de k minimisant la quantité ci-dessus. Des méthodes de ré-échantillonnage (jackknife, bootstrap) peuvent être mises en œuvre mais celles-ci sont coûteuses en temps de calcul. Une valeur heuristique de k peut être fixée en considérant le graphique des paramètres en fonction de k. Elle est choisie dans la zone où les valeurs absolues des paramètres commencent à se stabiliser.

5.3 Régression sur composantes principales

L'Analyse en Composantes Principales est, entre autre, la recherche de p variables dites principales qui sont des combinaisons linéaires des variables initiales de variance maximale sous une contrainte d'orthogonalité. En désignant par $\mathbf{V}$ la matrice des vecteurs propres de la matrice des corrélations $\mathbf{R}$ rangés dans l'ordre décroissant des valeurs propres, les valeurs prises par ces variables principales sont obtenues dans la matrice des composantes principales

$$\mathbf{C} = (\widetilde{\mathbf{X}} - \mathbf{1}\overline{\mathbf{x}}')\mathbf{V}.$$

Elles ont chacune pour variance la valeur propre λ_j associée. Le sous-espace engendré par ces variables principales est le même que celui engendré par les variables initiales. Il est donc géométriquement équivalent de régresser Y sur les colonnes de $\mathbf{C}$ que sur celles de $\widetilde{\mathbf{X}}$. Les problèmes de colinéarité sont alors résolus en supprimant les variables principales de plus faibles variances c'est-à-dire associées aux plus petites valeurs propres.

La solution obtenue présente ainsi de meilleures qualités prédictives mais, les coefficients de la régression s'appliquant aux composantes principales, un calcul complémentaire est nécessaire afin d'évaluer et d'interpréter les effets de chacune des variables initiales.

5.4 Modèles curvilinéaires

En cas d'invalidation de l'hypothèse de linéarité, il peut être intéressant de considérer des modèles polynômiaux, très classiques pour décrire des phénomènes physiques, de la forme

$$Y = \beta_0 + \cdots + \beta_j X^j + \cdots + \gamma_{kl} X^k X^l + \cdots + \delta_j X^{j2}$$

qui sont encore appelés *surfaces de réponse*. Ces modèles sont faciles à étudier dans le cadre linéaire, il suffit d'ajouter des nouvelles variables constituées des produits ou des carrés des variables explicatives initiales. Les choix : présence ou non d'une interaction entre deux variables, présence ou non d'un terme quadratique se traitent alors avec les mêmes outils que ceux des choix de variable mais en intégrant une contrainte lors de la lecture des résultats : ne pas considérer des modèles incluant des termes quadratiques dont les composants linéaires auraient été exclus ou encore, ne pas supprimer d'un modèle une variable d'un effet linéaire si elle intervient dans un terme quadratique.

La procédure `rsreg` de SAS est plus particulièrement adaptée aux modèles quadratiques. Elle ne comporte pas de procédure de choix de modèle mais fournit des aides et diagnostics sur l'ajustement de la surface ainsi que sur la recherche des points optimaux.

Attention : Ce type de modèle accroît considérablement les risques de colinéarité, il est peu recommandé de considérer des termes cubiques.

6 Influence, résidus, validation

Avant toute tentative de modélisation complexe, il est impératif d'avoir conduit des analyses uni et bivariées afin d'identifier des problèmes sur les distributions de chacune des variables : dissymétrie, valeurs atypiques (outliers) ou sur les liaisons des variables prises deux par deux : non-linéarité. Ces préliminaires acquis, des aides ou diagnostics associés à la régression linéaire multiple permettent de détecter des violations d'hypothèses (homoscédasticité, linéarité) ou des points influents dans ce contexte multidimensionnel.

6.1 Effet levier

Comme toute méthode quadratique, l'estimation des paramètres est très sensible à la présence de points extrêmes susceptibles de perturber gravement les résultats. À partir de l'équation de prédiction : $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ on remarque qu'une observation i est influente si le terme correspondant h_i^i de la diagonale de $\mathbf{H}$ est grand.

On écrit encore :

$$\mathbf{H} = \frac{\mathbf{1}\mathbf{1}'}{n} + \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'$$

et

$$h_i^i = \frac{1}{n} + (\mathbf{x}_i - \overline{\mathbf{x}})'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = \frac{1}{n} + \sum_{j=1}^{p}\left(\frac{\mathbf{v}'^j(\mathbf{x}_i - \overline{\mathbf{x}})}{\sqrt{\lambda_j}}\right)^2$$

où les $\lambda_j, \mathbf{v}^j$ sont respectivement les valeurs et vecteurs propres de la matrice $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$. Ainsi, plus une observation est éloignée du barycentre, et ce dans la direction d'un vecteur propre associé à une petite valeur propre, et plus cette observation a un effet levier important.

6.2 Résidus

Nous désignons comme précédemment par $\mathbf{b}_{(i)}, \widehat{\mathbf{y}}_{(i)}, \mathbf{e}_{(i)}$, et

$$s^2_{(i)} = \frac{\mathbf{e}'_{(i)}\mathbf{e}_{(i)}}{n - p - 2}$$

les estimations réalisées sans la $i$ème observation. Les expressions

$$\begin{aligned}
\mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{y}, \\
\mathbf{r} &= \text{diag}[s^2(1 - h_i^i)]^{-1/2}\mathbf{e}, \\
\mathbf{t} &= \text{diag}[s^2_{(i)}(1 - h_i^i)]^{-1/2}\mathbf{e}
\end{aligned}$$

définissent respectivement les résidus calculés, les résidus *standardisés* (chacun divisé par l'estimation de l'écart-type) et les résidus *studentisés* dans lesquels l'estimation de σ_u^2 ne fait pas intervenir la $i$ème observation.

De trop grands résidus sont aussi des signaux d'alerte. Par exemple, un résidu studentisé de valeur absolue plus grande que 2 peut révéler un problème.

6.3 Mesures d'influence

L'effet levier peut apparaître pour des observations dont les valeurs prises par les variables explicatives sont élevées (observation loin du barycentre $\overline{x}$). De grands résidus signalent plutôt des valeurs atypiques de la variable à expliquer. Les deux diagnostics précédents sont combinés dans des mesures synthétiques

proposées par différents auteurs. Les plus utilisées sont

$$D_i = \frac{1}{s^2(p+1)}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)}) = \left[\frac{h_i^i}{1-h_i^i}\right]\frac{r_i^2}{(p+1)}, \tag{2.1}$$

$$\text{DFFITS}_i = \frac{1}{s_{(i)}\sqrt{h_i^i}}(\widehat{y}_i - \widehat{y}_{(i)i}) = \left[\frac{h_i^i}{1-h_i^i}\right]^{1/2} t_i. \tag{2.2}$$

La première, notée *Cook's D* conclut à une influence de l'observation i lorsque la valeur de D_i dépasse 1.

D'autres mesures moins fréquemment utilisées sont proposées dans les logiciels. Certaines considèrent les écarts entre l'estimation d'un paramètre b_i et son estimation sans la $i$ème observation, une autre le rapport des déterminants des matrices de covariance des estimateurs des paramètres calculées avec et sans la $i$ème observation...

6.4 Régression partielle

Un modèle de régression multiple est une technique *linéaire*. Il est raisonnable de s'interroger sur la pertinence du caractère linéaire de la contribution d'une variable explicative à l'ajustement du modèle. Ceci peut être réalisé en considérant une *régression partielle*.

On calcule alors deux régressions :

- la régression de Y sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$, dans laquelle la $j$ème variable est omise, soit $\mathbf{r}_{y(j)}$ le vecteur des résidus obtenus.
- La régression de X^j sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$. Soit $\mathbf{r}_{x(j)}$ le vecteur des résidus obtenus.

La comparaison des résidus par un graphe (nuage de points $\mathbf{r}_{y(j)} \times \mathbf{r}_{x(j)}$) permet alors de représenter la nature de la liaison entre X^j et Y *conditionnellement* aux autres variables explicatives du modèle.

6.5 Graphes

Différents graphiques permettent finalement de contrôler le bien fondé des hypothèses de linéarité, d'homoscédasticité, éventuellement de normalité des résidus.

- Le premier considère le nuage de points des résidus studentisés croisés avec les valeurs prédites. Les points doivent être uniformément répartis entre les bornes -2 et $+2$ et ne pas présenter de formes suspectes.
- Le deuxième croise les valeurs observées de Y avec les valeurs prédites. Il illustre le coefficient de détermination R qui est aussi la corrélation linéaire simple entre $\widehat{\mathbf{y}}$ et $\mathbf{y}$. Les points doivent s'aligner autour de la première bissectrice. Il peut être complété par l'intervalle de confiance des y_i ou celui de leurs moyennes.
- La qualité, en terme de linéarité, de l'apport de chaque variable est étudiée par des régressions partielles. Chaque graphe de résidus peut être complété par une estimation fonctionnelle ou régression non-paramétrique (loess, noyau, spline) afin d'en facilité la lecture.
- Le dernier trace la droite de Henri (Normal QQplot) des résidus dont le caractère linéaire de la représentation donne une idée de la normalité de la distribution.

7 Exemple

7.1 Les données

Elles sont extraites de Jobson (1991) et décrivent les résultats comptables de 40 entreprises du Royaume Uni.

Descriptif des 13 variables (en anglais pour éviter des traductions erronées) :

RETCAP	Return on capital employed
WCFTDT	Ratio of working capital flow to total debt
LOGSALE	Log to base 10 of total sales
LOGASST	Log to base 10 of total assets
CURRAT	Current ratio
QUIKRAT	Quick ratio
NFATAST	Ratio of net fixed assets to total assets
FATTOT	Gross sixed assets to total assets
PAYOUT	Payout ratio
WCFTCL	Ratio of working capital flow to total current liabilities
GEARRAT	Gearing ratio (debt-equity ratio)
CAPINT	Capital intensity (ratio of total sales to total assets)
INVTAST	Ratio of total inventories to total assets

7.2 Résultat du modèle complet

La procédure SAS/REG est utilisée dans le programme suivant. La plupart des options sont actives afin de fournir la plupart des résultats même si certains sont redondants ou peu utiles.

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp1 all;
 model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       /dw covb  Influence cli clm tol vif collin R P;
output out=resout h=lev p=pred r=res student=resstu ;
run;
```

Les résultats ne sont pas listés de façon exhaustive, les matrices et tableaux trop volumineux et peu significatifs ont été tronqués.

```
Descriptive Statistics
Variables       Sum           Mean      Uncorrected SS        Variance       Std Deviation
INTERCEP        40             1            40                  0                 0
WCFTCL          10.29        0.25725         6.4339        0.0970973718       0.3116045118
WCFTDT           9.04        0.226           4.9052        0.0733887179       0.2709035215
...
CURRAT          72.41        1.81025       279.0039        3.7929153205       1.9475408392
RETCAP           5.71        0.14275         1.5233        0.0181589103       0.1347550009

Uncorrected Sums of squares and Crossproducts
USSCP      INTERCEP     WCFTCL      WCFTDT     GEARRAT     LOGSALE     LOGASST     NFATAST
INTERCEP       40       10.29        9.04        12.2       173.7       174.81       13.46
WCFTCL      10.29      6.4339      5.4926      1.5997     40.8722     46.2433      3.5523
WCFTDT       9.04      5.4926      4.9052      1.3972     34.4091     39.8937      2.9568
...
CURRAT      72.41      35.222      33.248     16.3188    265.2051     314.449     20.4126
RETCAP       5.71      2.0009      1.6226      1.5391     26.3636      25.379      1.6199

Correlation
CORR         WCFTCL      WCFTDT     GEARRAT     LOGSALE     LOGASST     NFATAST     CAPINT
WCFTCL       1.0000      0.9620     -0.5520     -0.3100      0.1829      0.0383     -0.2376
WCFTDT       0.9620      1.0000     -0.5611     -0.4533      0.0639     -0.0418     -0.2516
GEARRAT     -0.5520     -0.5611      1.0000      0.2502      0.0387     -0.0668      0.2532
...
CURRAT       0.7011      0.8205     -0.3309     -0.6406     -0.0460     -0.2698     -0.3530
RETCAP       0.3249      0.2333     -0.1679      0.2948      0.1411     -0.2974      0.3096
```

La matrice des corrélations montre des valeurs élevées, on peut déja s'attendre à des problèmes de colinéarité.

```
Model Crossproducts X'X X'Y Y'Y
X'X        INTERCEP     WCFTCL     WCFTDT    GEARRAT    LOGSALE    LOGASST    NFATAST
INTERCEP         40      10.29       9.04       12.2      173.7     174.81      13.46
WCFTCL        10.29     6.4339     5.4926     1.5997    40.8722    46.2433     3.5523
WCFTDT         9.04     5.4926     4.9052     1.3972    34.4091    39.8937     2.9568
...
X'X Inverse, Parameter Estimates, and SSE
          INTERCEP     WCFTCL     WCFTDT    GEARRAT    LOGSALE    LOGASST    NFATAST
INTERCEP  3.2385537  1.3028641  -1.570579  -0.05877  0.3001809  -0.826512  -0.238509
WCFTCL    1.3028641  7.0714100  -9.955073  -0.54391  -0.007877  -0.292412  -0.233915
WCFTDT    -1.570579  -9.955073  15.968504  1.582975  0.0112826  0.3138925   0.149976
...
Analysis of Variance
                          Sum of          Mean
Source          DF       Squares        Square         F Value       Prob>F
                (1)
Model           12      0.55868 (2)     0.04656 (5)    8.408 (7)     0.0001 (8)
Error           27      0.14951 (3)     0.00554 (6)
C Total         39      0.70820 (4)
  Root MSE       0.07441 (9)    R-square      0.7889 (12)
  Dep Mean       0.14275 (10)    Adj R-sq     0.6951 (13)
  C.V.          52.12940 (11)
```

(1)	degrés de liberté de la loi de Fisher du test global
(2)	SSR
(3)	SSE ou déviance
(4)	SST=SSE+SSR
(5)	SSR/DF
(6)	s^2 =MSE=SSE/DF est l'estimation de σ_u^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10)
(12)	Coefficient de détermination R^2
(13)	Coefficient de détermination ajusté R'^2

```
Parameter Estimates
                    Parameter      Standard    T for H0:                        Variance
Variable  DF       Estimate         Error   Parameter=0 Prob>|T|   Tolerance    Inflation
                      (1)            (2)        (3)        (4)         (5)         (6)
INTERCEP   1      0.188072     0.13391661      1.404     0.1716    .            0.00000000
WCFTCL     1      0.215130     0.19788455      1.087     0.2866    0.03734409  26.77799793
WCFTDT     1      0.305557     0.29736579      1.028     0.3133    0.02187972  45.70441500
GEARRAT    1     -0.040436     0.07677092     -0.527     0.6027    0.45778579   2.18442778
LOGSALE    1      0.118440     0.03611612      3.279     0.0029    0.10629382   9.40788501
LOGASST    1     -0.076960     0.04517414     -1.704     0.0999    0.21200778   4.71680805
NFATAST    1     -0.369977     0.13739742     -2.693     0.0120    0.20214372   4.94697537
CAPINT     1     -0.014138     0.02338316     -0.605     0.5505    0.37587215   2.66047911
FATTOT     1     -0.100986     0.08764238     -1.152     0.2593    0.23929677   4.17891139
INVTAST    1      0.250562     0.18586858      1.348     0.1888    0.13770716   7.26178633
PAYOUT     1     -0.018839     0.01769456     -1.065     0.2965    0.84271960   1.18663431
QUIKRAT    1      0.176709     0.09162882      1.929     0.0644    0.00408524 244.78377222
CURRAT     1     -0.223281     0.08773480     -2.545     0.0170    0.00486336 205.61923071
```

(1)	estimations des paramètres (b_j)
(2)	écarts-types de ces estimations (s_{b_j})
(3)	statistique T du test de Student de $H_0 : b_j = 0$
(4)	$P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$
(5)	$1 - R^2_{(j)}$
(6)	VIF=$1/(1 - R^2_{(j)})$

Ces résultats soulignent les problèmes de colinéarités. De grands “VIF” sont associés à de grands écart-types des estimations des paramètres. D’autre part les nombreux tests de Student non significatifs renforcent l’idée que trop de variables sont présentes dans le modèle.

```
Covariance of Estimates
COVB       INTERCEP     WCFTCL     WCFTDT    GEARRAT    LOGSALE    LOGASST    NFATAST
INTERCEP  0.0179336  0.0072146  -0.008697  -0.000325  0.0016622  -0.004576  -0.001320
WCFTCL    0.0072146  0.039158   -0.055126  -0.003011  -0.000043   -0.00161   -0.00129
WCFTDT    -0.008697  -0.055126  0.0884264  0.0087658  0.0000624  0.0017381  0.0008305
...

Collinearity Diagnostics
                  Condition
  Eigenvalue        Index

     8.76623     1.00000
     2.22300     1.98580
     0.68583     3.57518
     0.56330     3.94489
     0.31680     5.26036
     0.18140     6.95173
     0.12716     8.30291
     0.08451    10.18479
     0.02761    17.82007
     0.01338    25.59712
     0.00730    34.66338
     0.00223    62.63682
     0.00125    83.83978
```

Valeurs propres de $\mathbf{X}'\mathbf{X}$ et indice de conditionnement égal au rapport $\sqrt{\lambda_1/\lambda_j}$. Les grandes valeurs (> 10) insistent encore sur le mauvais conditionnement de la matrice à inverser.

```
    Dep Var   Predict  Std Err Lower95   Upper95  Lower95   Upper95          Std Err  Student
Obs  RETCAP    Value   Predict   Mean      Mean    Predict   Predict  Residual Residual Residual
       (1)      (2)      (3)      (4)       (5)       (6)       (7)      (8)     (9)      (10)
1    0.2600    0.2716  0.053    0.1625    0.3808    0.0839    0.4593  -0.0116   0.052   -0.223
2    0.5700    0.3690  0.039    0.2882    0.4497    0.1962    0.5417   0.2010   0.063    3.183
3    0.0900   0.00897  0.063   -0.1205    0.1385   -0.1912    0.2092   0.0810   0.039    2.055
4    0.3200    0.2335  0.021    0.1903    0.2768    0.0748    0.3922   0.0865   0.071    1.212
5    0.1700    0.1164  0.046    0.0215    0.2113   -0.0634    0.2961   0.0536   0.058    0.920
6    0.2400    0.2542  0.033    0.1864    0.3219    0.0871    0.4212  -0.0142   0.067   -0.213
...
                       Cook's            Hat Diag    Cov                INTERCEP    WCFTCL    WCFTDT
Obs   -2-1-0 1 2         D     Rstudent      H       Ratio    Dffits    Dfbetas    Dfbetas   Dfbetas
         (11)          (12)     (13)       (14)      (15)     (15)      (15)       (15)      (15)
 1  |      |       | 0.004   -0.2194    0.5109    3.2603   -0.2242    0.0299    0.0632 -0.0911
 2  |      |****** | 0.302    3.9515    0.2795    0.0050    2.4611    0.9316   -0.3621  0.3705
 3  |      |****   | 0.832    2.1955    0.7192    0.6375    3.5134    0.5543    2.1916 -2.0241
 4  |      |**     | 0.010    1.2228    0.0803    0.8585    0.3613   -0.0132   -0.0835  0.1207
 5  |      |*      | 0.041    0.9175    0.3864    1.7591    0.7280   -0.0386    0.0906  0.0060
 6  |      |       | 0.001   -0.2088    0.1969    1.9898   -0.1034    0.0189   -0.0203  0.0243
15  |   ***|       | 0.150   -1.9223    0.3666    0.4583   -1.4623   -0.2063    0.3056 -0.6231
16  |      |***    | 3.471    1.6394    0.9469    8.5643    6.9237   -0.9398    0.2393 -0.2323
17  |      |       | 0.000    0.1401    0.1264    1.8514    0.0533    0.0223    0.0090 -0.0113
20  |   ***|       | 0.054   -1.9588    0.1677    0.3278   -0.8794   -0.0360   -0.3302  0.4076
21  |  ****|       | 4.970   -2.2389    0.9367    2.6093   -8.6143   -1.2162    0.1768 -0.1422
       ...
```

(1)	variable à expliquer y_i
(2)	valeur ajustée $\widehat{y}_i$
(3)	écart-type de cette estimation$s_{\widehat{y}_i}$
(4)et (5)	Intervalle de confiance pour l'estimation de $E(y_i)$
(6) et (7)	Intervalle de confiance pour l'estimation de y_i
(8)	résidus calculés e_i
(9)	écarts-types de ces estimations
(10)	résidus standardisés (ou studentisés internes) r_i
(11)	repérage graphique des résidus standardisés : $* = 0.5$.
(12)	Distance de Cook
(13)	résidus studentisés (externes) t_i
(14)	Termes diagonaux de la matrice chapeau **H**
(15)	autres indicateurs d'influence

Seules les observations 16 et 21 seraient à inspecter avec attention.

```
Sum of Residuals                     0
Sum of Squared Residuals        0.1495  (SSE)
Predicted Resid SS (Press)      1.0190  (PRESS)
```

Sélection du modèle

Parmi les trois types d'algorithmes et les différents critères de choix, une des façons les plus efficaces consistent à choisir les options du programme ci-dessous. Tous les modèles (parmi les plus intéressants selon l'algorithme de Furnival et Wilson) sont considérés. Seul le meilleur pour chaque niveau, c'est-à-dire pour chaque valeur p du nombre de variables explicatives sont donnés. Il est alors facile de choisir celui minimisant l'un des critères globaux (C_p ou BIC ou ...).

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp2 ;
model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       / selection=rsquare cp rsquare bic best=1;
run;
```

```
   N = 40     Regression Models for Dependent Variable: RETCAP
  R-square Adjusted  C(p)     BIC    Variables in Model
In         R-square
 1  0.1055 0.0819 78.3930 -163.26 WCFTCL
 2  0.3406 0.3050 50.3232 -173.72 WCFTDT QUIKRAT
 3  0.6154 0.5833 17.1815 -191.14 WCFTCL NFATAST CURRAT
 4  0.7207 0.6888  5.7146 -199.20 WCFTDT LOGSALE NFATAST CURRAT
 5  0.7317 0.6923  6.3047 -198.05 WCFTDT LOGSALE NFATAST QUIKRAT CURRAT
 6  0.7483 0.7025  6.1878 -197.25 WCFTDT LOGSALE NFATAST INVTAST QUIKRAT CURRAT
 7  0.7600 0.7075  6.6916 -195.77 WCFTDT LOGSALE LOGASST NFATAST FATTOT QUIKRAT CURRAT
 8  0.7692 0.7097  7.5072 -193.87 WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT CURRAT
 9  0.7760 0.7088  8.6415 -191.59 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT
                                   CURRAT
10  0.7830 0.7082  9.7448 -189.15 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST PAYOUT
                                   QUIKRAT CURRAT
11  0.7867 0.7029 11.2774 -186.40 WCFTCL WCFTDT LOGSALE LOGASST NFATAST CAPINT FATTOT INVTAST
                                   PAYOUT QUIKRAT CURRAT
12  0.7888 0.6950 13.0000 -183.51 WCFTCL WCFTDT GEARRAT LOGSALE LOGASST NFATAST CAPINT FATTOT
                                   INVTAST PAYOUT QUIKRAT CURRAT
```

Dans cet example, C_p et BIC se comportent de la même façon. Avec peu de variables, le modèle est trop biaisé. Ils atteignent un minimum pour un modèle à 4 variables explicatives puis croissent de nouveau selon la première bissectrice. La maximisation du R^2 ajusté conduirait à une solution beaucoup moins parcimonieuse. On note par ailleurs que l'algorithme remplace WCFTCL par WCFTDT. Un algorithme par sélection ne peut pas aboutir à la solution optimale retenue.

Résultats du modèle réduit

```
proc reg data=sasuser.ukcomp1 all;
model RETCAP =  WCFTDT  NFATAST   LOGSALE  CURRAT
       /dw Influence cli clm tol vif collin r p ;
output out=resout h=lev p=pred r=res student=resstu ;
plot (student. r.)*p.;
plot p.*retcap;
run;
```

```
Analysis of Variance
                          Sum of          Mean
Source           DF      Squares        Square      F Value       Prob>F
Model             4      0.51043       0.12761       22.583       0.0001
Error            35      0.19777       0.00565
C Total          39      0.70820

       Root MSE       0.07517     R-square       0.7207
       Dep Mean       0.14275     Adj R-sq       0.6888
       C.V.          52.65889

Parameter Estimates
              Parameter       Standard    T for H0:                                Variance
Variable  DF  Estimate          Error    Parameter=0    Prob > |T|  Tolerance     Inflation
INTERCEP   1  0.024204    0.07970848       0.304        0.7632     .            0.00000000
WCFTDT     1  0.611885    0.08257125       7.410        0.0001    0.28956358    3.45347296
NFATAST    1 -0.474448    0.07015433      -6.763        0.0001    0.79119995    1.26390301
LOGSALE    1  0.060962    0.01606877       3.794        0.0006    0.54792736    1.82505944
CURRAT     1 -0.068949    0.01321091      -5.219        0.0001    0.21887292    4.56886122

Collinearity Diagnostics
                      Condition Var Prop Var Prop  Var Prop  Var Prop  Var Prop
Number  Eigenvalue      Index  INTERCEP  WCFTDT    NFATAST   LOGSALE   CURRAT
     1     3.86169     1.00000  0.0014   0.0076    0.0098    0.0016    0.0052
     2     0.87647     2.09904  0.0014   0.0608    0.0355    0.0046    0.0427
     3     0.17128     4.74821  0.0206   0.1731    0.5177    0.0170    0.0667
     4     0.07821     7.02670  0.0026   0.7201    0.4369    0.0388    0.5481
     5     0.01235    17.68485  0.9741   0.0384    0.0000    0.9381    0.3373

                     Cook's          Hat Diag  Cov                INTERCEP  WCFTDT  NFATAST
Obs   -2-1-0 1 2         D  Rstudent    H     Ratio  Dffits   Dfbetas Dfbetas  Dfbetas
15  |   ***|      |  0.211  -1.9115  0.2372   0.9096 -1.0659 -0.0240  -0.8161  -0.3075
16  |      |*     |  1.554   0.9919  0.8876   8.9162  2.7871  0.0320  -0.0746   0.1469
17  |      |      |  0.001   0.3866  0.0460   1.1854  0.0849  0.0348  -0.0430   0.0256

Sum of Residuals            0
Sum of Squared Residuals    0.1978  (Par rapport au modèle complet, la déviance augmente
Predicted Resid SS (Press)  0.3529   mais PRESS diminue très sensiblement)
```

8 Exercices

Exo 1

Nous supposons vérifiées les hypothèses relatives au modèle de régression linéaire multiple pour les observations $(y_i, x_i^1, \ldots, x_i^n)$ des variables statistiques $Y, X^1, \ldots, X^p$.

i. Calculer les moments (espérance et variance) des estimateurs $\mathbf{b}, \widehat{\mathbf{y}}$ et $\mathbf{e}$ de $\boldsymbol{\beta}, \mathbf{y}$ et $\mathbf{u}$. Calculer $E(\mathbf{e}'\mathbf{e})$.

ii. Montrer que

$$(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \quad (2.3)$$

$$\widehat{\mathbf{y}}'\widehat{\mathbf{y}} = \mathbf{y}'\mathbf{H}\mathbf{y} \quad (2.4)$$

$$\mathbf{y}'\mathbf{y} = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} + (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}). \quad (2.5)$$

En déduire que : SST=SSE+SSR.

Exo 2

Pour simplifier les calculs, on suppose dans cet exercice que les variables sont centrées ($\overline{\mathbf{x}} = 0, \overline{y} = 0$) et on s'intéresse à la régression sans terme constant ($\beta_0 = 0$). On admettra le résultat suivant du calcul par bloc de l'inverse $\mathbf{B}$ d'une matrice carrée régulière $\mathbf{A}$:

$$\mathbf{B}_{11} = [\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1}.$$

i. Montrer que le coefficient de détermination vérifie : $r^2 = \frac{\|\widehat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2}$.

ii. Soit $\mathbf{r}_1$ le vecteur contenant les coefficients de corrélation linéaire empirique entre X^1 et chacune des variables $X^2, \ldots, X^p$, $\mathbf{R}_{(1)}$ la matrice des corrélations des $X^2, \ldots, X^p$ et $\mathbf{R}$ la matrice de corrélations de toutes les variables X^j. On note également $r^2_{(1)}$ le coefficient de détermination de la régression de la variable X^1 sur les variables $X^2, \ldots, X^p$. Montrer que $r^2_{(1)} = \mathbf{r}'_1\mathbf{R}^{-1}_{(1)}\mathbf{r}_1$.

iii. En déduire que $[\mathbf{R}^{-1}]^1_1 = \frac{1}{1-r^2_{(1)}}$.

iv. Commentaire pour les autres variables et l'indicateur de colinéarité.

Exo 3

On reprend les notations usuelles de la régression linéaire multiple et on désigne par $\mathbf{x}_n$ la dernière ligne de $\mathbf{X}$ et par $\mathbf{X}_{(n)}$ la matrice $(n-1) \times (p+1)$ privée de cette dernière ligne.

i. Montrer que $\mathbf{X}'\mathbf{X} = \mathbf{X}'_{(n)}\mathbf{X}_{(n)} + \mathbf{x}_n\mathbf{x}'_n$.

ii. Soit $\mathbf{A}$ une matrice symétrique régulière et $\mathbf{b}, \mathbf{c}$ deux vecteurs à $(p+1)$ composantes. Montrer que l'inverse de la matrice $\mathbf{A} + \mathbf{b}\mathbf{c}'$ est la matrice $\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{c}'\mathbf{A}^{-1}}{1+\mathbf{b}'\mathbf{A}^{-1}\mathbf{c}}$

iii. Trouver l'expression de h_{nn} dans la décomposition suivante :

$$[\mathbf{X}'_{(n)}\mathbf{X}_{(n)}]^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{1-h_{nn}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_n\mathbf{x}'_n(\mathbf{X}'\mathbf{X})^{-1}.$$

iv. Montrer que $\mathbf{X}_{(n)}\mathbf{y}_{(n)} = \mathbf{X}'\mathbf{y} - \mathbf{x}_n y_n$. Montrer ensuite que

$$\mathbf{b}_{(n)} = \mathbf{b} - \frac{1}{1-h_{nn}}(X'X)^{-1}\mathbf{x}_n(y_n - \mathbf{x}'_n\mathbf{b}).$$

Discuter de l'impact sur $\mathbf{b}$ de la suppression de l'observation n.

v. Montrer que la distance de Cook

$$D_n = \frac{1}{(p+1)s^2}(\mathbf{b}_{(n)} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_{(n)} - \mathbf{b})$$

se met sous la forme :

$$D_n = \frac{h_{nn}}{1-h_{nn}}\frac{e_n^2}{(p+1)s^2(1-h_{nn})}.$$

Exo 4

L'objet de cet exercice est de construire un indicateur permettant de comparer des modèles pour leurs qualités prédictives. On considère un premier modèle complet (avec toutes les variables) supposé *vrai* :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \in \text{Vect}(\mathbf{X}^1, \ldots, \mathbf{X}^p), \quad \text{rang}(\mathbf{X}) = p, \quad \mathbf{u} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n). \tag{2.6}$$

Un deuxième modèle est un sous-modèle du précédent et donc légèrement faux. La matrice $\mathbf{Z}$ de ce modèle est supposée de plein rang $(q+1) < (p+1)$ et contient donc un sous-ensemble des colonnes de $\mathbf{X}$. Ainsi, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ n'appartient pas nécessairement à l'espace vectoriel engendré par les colonnes de $\mathbf{Z}$. On note $\boldsymbol{\alpha}_0$ les paramètres les moins mauvais pour le 2ème modèle. Ils sont obtenus par la projection de $\mathbf{X}\boldsymbol{\beta}$ sur $\text{Vect}(\mathbf{Z}^1, \ldots, \mathbf{Z}^q)$:

$$\mathbf{Z}\boldsymbol{\alpha}_0 = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}.$$

On note enfin $\mathbf{a}$ les paramètres du 2ème modèle estimés par les moindres carrés.

i. Montrer que $E(\mathbf{a}) = \boldsymbol{\alpha}_0$.

ii. On note $\widehat{\mathbf{y}} = \mathbf{Za}$ la prévision de $\mathbf{y}$ par le 2ème modèle. Montrer que trace(Var($\widehat{\mathbf{y}}$))$= \sigma^2(q+1)$.

iii. Soit $E_p = E\|\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}\|^2$ l'erreur quadratique moyenne de prédiction pour le 2ème modèle. Montrer que

$$E_p = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}_0\|^2 + \sigma^2(q+1)$$

qui se décompose donc en le carré du biais plus la variance.
Suggestion : Calculer $E\|\widehat{\mathbf{y}} - \mathbf{Z}\boldsymbol{\alpha}_0 + \mathbf{Z}\boldsymbol{\alpha}_0 - \mathbf{X}\boldsymbol{\beta}\|^2$.

iv. Soit $\text{SSE}_q = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2$ la somme des carrés des résidus du 2ème modèle. Montrer que

$$E(\text{SSE}_q) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}_0\|^2 + \sigma^2(n-q-1).$$

Suggestion : Noter que $\mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \widehat{\mathbf{y}} - E(\mathbf{y} - \widehat{\mathbf{y}}) + E(\mathbf{y} - \widehat{\mathbf{y}})$.

v. Le problème est d'estimer E_p. On estime sans biais σ^2 par $s^2 = \text{SSE}/(n-p-1)$. Montrer que

$$\widehat{E}_p = \text{SSE}_q - (n-2q-2)s^2$$

est un estimateur sans biais de E_p c'est-à-dire que $E(\widehat{E}_p) = E_p$.

vi. Le C_p de Mallows est une version standardisée de l'erreur de prévision : $C_p = \frac{\text{SSE}_q}{s^2} - (n-2q-2)$. Dans l'hypothèse où le sous-modèle est exact : $\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\alpha}_0$, montrer qu'alors le (C_p) est "proche" de $q+1$. On acceptera pour "meilleur" un modèle biaisé à condition qu'il induise une baisse significative de la variance et ainsi de l'erreur quadratique moyenne de prévision.

vii. Dans l'exemple ci-dessous, calculer le C_p de Mallows du sous-modèle.

```
modele pert = Tlumin lumin Txgn xgn Txy xy xa xb
   Model            8        1007.62105    125.95263      229.845       0.0001
   Error          180          98.63792      0.54799
   C Total        188        1106.25897

modele pert = lumin Txgn xy xa xb
   Model            5       1007.11132    201.42226      371.772       0.0001
   Error          183         99.14764      0.54179
   C Total        188       1106.25897
```

Chapitre 3

Analyses de variance et covariance

1 Introduction

Les techniques dites d'*analyse de variance* sont des outils entrant dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par une ou plusieurs variables qualitatives. L'objectif essentiel est alors de comparer les moyennes empiriques de la variable quantitative observées pour différentes catégories d'unités statistiques. Ces catégories sont définies par l'observation des variables qualitatives ou *facteurs* prenant différentes modalités ou encore de variables quantitatives découpées en classes ou *niveaux*. Une combinaison de niveaux définit une *cellule*, *groupe* ou *traitement*.

Il s'agit donc de savoir si un facteur ou une combinaison de facteurs (*interaction*) a un *effet* sur la variable quantitative en vue, par exemple, de déterminer des conditions optimales de production ou de fabrication, une dose optimale de médicaments.... Ces techniques apparaissent aussi comme des cas particuliers de la régression linéaire multiple en associant à chaque modalité une *variable indicatrice* (dummy variable) et en cherchant à expliquer une variable quantitative par ces variables indicatrices. L'appellation "analyse de variance" vient de ce que les tests statistiques sont bâtis sur des comparaisons de sommes de carrés de variations.

L'analyse de variance est souvent utilisée pour analyser des données issue d'une *planification d'expérience* au cours de laquelle l'expérimentateur a la possibilité de contrôler *a priori* les niveaux des facteurs avec pour objectif d'obtenir le maximum de précision au moindre coût. Ceci conduit en particulier à construire des facteurs orthogonaux deux à deux (variables explicatives non linéairement corrélées) afin de minimiser la variance des estimateurs (cf. chapitre précédent). On distingue le cas particulier important où les cellules ont le même effectif, on parle alors de *plan orthogonal* ou *équirépété* ou *équilibré* (balanced), qui conduit à des simplifications importantes de l'analyse de variance associée. On appelle plan *complet* un dispositif dans lequel toutes les combinaisons de niveaux ont été expérimentées. On distingue entre des modèles fixes, aléatoires ou mixtes selon le caractère déterministe (contrôlé) ou non des facteurs par exemple si les modalités résultent d'un choix aléatoire parmi un grand nombre de possibles. Seuls les modèles fixes sont considérés.

L'analyse de covariance considère une situation plus générale dans laquelle les variables explicatives sont à la fois quantitatives, appelées covariables, et qualitatives ou facteurs. L'objectif est alors de comparer, non plus des moyennes par cellules, mais les paramètres des différents modèles de régressions estimées pour chaque combinaison de niveau. Ce type de modèle est introduit en fin de chapitre.

Les spécificités de la planification d'expérience ne seront qu'abordées dans ce chapitre. Les applications en sont surtout développées en milieu industriel : contrôle de qualité, optimisation des processus de production, ou en agronomie pour la sélection de variétés, la comparaison d'engrais, d'insecticides.... La bibliographie est abondante à ce sujet.

2 Modèle à un facteur

Cette situation est un cas particulier d'étude de relations entre deux variables statistiques : une quantitative Y admettant une densité et une qualitative T ou facteur qui engendre une partition ou classification de l'échantillon en J groupes, cellules ou classes indicées par j. L'objectif est de comparer les distributions de Y pour chacune des classes en particulier les valeurs des moyennes et variances.

Un préalable descriptif consiste à réaliser un graphique constitué de boites à moustaches parallèles : une pour chaque modalité. Cette représentation donne une première appréciation de la comparaison des distributions (moyenne, variance) internes à chaque groupe.

2.1 Modèles

Pour chaque niveau j de T, on observe n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de la variable Y et où $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon ($n > J$). On suppose qu'à l'intérieur de chaque cellule, les observations sont indépendantes équidistribuées de moyenne μ_j et de variance *homogène* $\sigma_j^2 = \sigma^2$. Ceci s'écrit :

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests. Cette dernière hypothèse n'étant pas la plus sensible. Les espérances μ_j ainsi que le paramètre de nuisance σ^2 sont les paramètres inconnus à estimer.

On note respectivement :

$$\begin{aligned}
\bar{y}_{.j} &= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \\
s_j^2 &= \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2, \\
\bar{y}_{..} &= \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{J} y_{ij},
\end{aligned}$$

les moyennes et variances empiriques de chaque cellule, la moyenne générale de l'échantillon.

Les paramètres μ_j sont estimés sans biais par les moyennes $\bar{y}_{.j}$ et comme le modèle s'écrit alors :

$$y_{ij} = \bar{y}_{.j} + (y_{ij} - \bar{y}_{.j}),$$

l'estimation des erreurs est $e_{ij} = (y_{ij} - \bar{y}_{.j})$ tandis que les valeurs prédites sont $\widehat{y}_{ij} = \bar{y}_{.j}$.

Sous l'hypothèse d'homogénéité des variances, la meilleure estimation sans biais de σ^2 est

$$s^2 = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2}{n - J} = \frac{1}{n - J} [(n-1)s_1^2 + \cdots + (n_J - 1)s_J^2]$$

qui s'écrit donc comme une moyenne pondérée des variances empiriques de chaque groupe.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij} | i = 1, n_j; j = 1, J]'$ mis en colonne, $\mathbf{u} = [\varepsilon_{ij} | i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. Le $i$ème élément d'une variable indicatrice (dummy variable) $\mathbf{1}_j$ prend la valeur 1 si la $i$ème observation y_i est associée au $j$ème et 0 sinon.

Comme dans le cas de la régression linéaire multiple, le modèle consiste à écrire que l'espérance de la variable Y appartient au sous-espace linéaire engendré par les variables explicatives, ici les variables indicatrices :

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{1}_1 + \cdots + \beta_J \mathbf{1}_J + \mathbf{u}.$$

La matrice $\mathbf{X}$ alors construite n'est pas de plein rang $p + 1$ mais de rang p. La matrice $\mathbf{X}'\mathbf{X}$ n'est pas inversible et le modèle admet une infinité de solutions. Nous disons que les paramètres β_j ne sont pas

estimables ou identifiables. En revanche, certaines fonctions (combinaisons linéaires) de ces paramètres sont estimables et appelées *contrastes*.

Dans le cas du modèle d'analyse de variance à *un* facteur, la solution la plus simple adoptée consiste à considérer un sous-ensemble des indicatrices ou de combinaisons des indicatrices de façon à aboutir à une matrice inversible. Ceci conduit à considérer différents modèles associés à différentes paramétrisation. *Attention*, les paramètres β_j ainsi que la matrice $\mathbf{X}$ prennent à chaque fois des significations différentes.

Un premier modèle (cell means model) s'écrit comme celui d'une régression linéaire multiple sans terme constant avec $\boldsymbol{\beta} = [\mu_1, \ldots, \mu_J]'$ le vecteur des paramètres :

$$\begin{aligned} \mathbf{y} &= \beta_1 \mathbf{1}_1 + \cdots + \beta_J \mathbf{1}_J + \mathbf{u} \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \end{aligned}$$

Les calculs se présentent simplement (cf. exo 1) mais les tests découlant de ce modèle conduiraient à étudier la nullité des paramètres alors que nous sommes intéressés par tester l'égalité des moyennes.

Une autre paramétrisation, considérant cette fois le vecteur $\boldsymbol{\beta} = [\mu_J, \mu_1 - \mu_J, \ldots, \mu_{J-1} - \mu_J]'$ conduit à écrire le modèle (base cell model) de régression avec terme constant :

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{1}_1 + \cdots + \beta_{J-1} \mathbf{1}_{J-1} + \mathbf{u}.$$

C'est celle de SAS alors que Systat considère des paramètres d'effet différentiel $\mu_j - \mu_.$ par rapport à l'effet moyen $\mu_. = 1/J \sum_{j=1}^{J} \mu_j$. Ce dernier est encore un modèle (group effect model) de régression linéaire avec terme constant mais dont les variables explicatives sont des différences d'indicatrices et avec $\boldsymbol{\beta} = [\mu_., \mu_1 - \mu_., \ldots, \mu_{J-1} - \mu_.]'$:

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 (\mathbf{1}_1 - \mathbf{1}_J) + \cdots + \beta_{J-1} (\mathbf{1}_{J-1} - \mathbf{1}_J) + \mathbf{u}.$$

2.2 Test

On désigne les différentes sommes des carrés des variations par :

$$\begin{aligned} \text{SST} &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}^2 - n\bar{y}_{..}^2, \\ \text{SSW} &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2 = \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{J} n_j \bar{y}_{.j}^2, \\ \text{SSB} &= \sum_{j=1}^{J} n_j (\bar{y}_{.j} - \bar{y}_{..})^2 = \sum_{j=1}^{J} n_j \bar{y}_{.j}^2 - n\bar{y}_{..}^2, \end{aligned}$$

où "T" signifie totale, "W" (within) intra ou résiduelle, "B" (between) inter ou expliquée par la partition. Il est facile de vérifier que SST=SSB+SSW.

On considère alors l'hypothèse

$$H_0 \ : \mu_1 = \cdots = \mu_J,$$

qui revient à dire que la moyenne est indépendante du niveau ou encore que le facteur n'a pas d'effet, contre l'hypothèse

$$H_1 \ : \exists (j,k) \text{ tel que } \mu_j \neq \mu_k$$

qui revient à reconnaître un effet ou une influence du facteur sur la variable Y.

Dans les modèles précédents, l'étude de cette hypothèse revient à comparer par un test de Fisher un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres β_j et donc l'égalité des moyennes à celle de la dernière cellule ou à la moyenne générale.

Les résultats nécessaires à la construction du test qui en découle sont résumés dans la table d'analyse de la variance :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	F
Modèle (inter)	$J-1$	SSB	MSB=SSB/$(J-1)$	MSB/MSW
Erreur (intra)	$n-J$	SSW	MSW=SSW/$(n-J)$	
Total	$n-1$	SST		

Pratiquement, un programme de régression usuel permet de construire estimation et test de la nullité des β_j sauf pour le premier modèle qui doit tester l'égalité au lieu de la nullité des paramètres.

Dans le cas de deux classes $(J=2)$ on retrouve un test équivalent au test de Student de comparaison des moyennes de deux échantillons indépendants.

2.3 Comparaisons multiples

Si l'hypothèse nulle est rejetée, la question suivante consiste à rechercher quelles sont les groupes ou cellules qui possèdent des moyennes significativement différentes. De nombreux tests et procédures ont été proposés dans la littérature pour répondre à cette question.

Une procédure naïve consiste à exprimer, pour chaque paire j et l de groupes, un intervalle de confiance au niveau $100(1-\alpha)\%$ de la différence $(\mu_j - \mu_l)$:

$$(\bar{y}_{.j} - \bar{y}_{.l}) \pm t_{\alpha/2;(n-J)} s \left[\frac{1}{n_j} + \frac{1}{n_l}\right]^{1/2}.$$

Si, pour un couple (j,l) fixé a priori, cet intervalle inclut 0, les moyennes ne sont pas jugées significativement différentes au niveau α. L'orthogonalité des facteurs rendant les tests indépendants justifierait cette procédure mais elle ne peut être systématisée. En effet, si J est grand, il y a un total de $J(J-1)/2$ comparaisons à considérer et on peut s'attendre à ce que, sur le simple fait du hasard, $0,05 \times J(J-1)/2$ paires de moyennes soient jugées significativement différentes même si le test global accepte l'égalité des moyennes.

D'autres procédures visent à corriger cette démarche afin de contrôler globalement le niveau des comparaisons. Certaines proposent des intervalles plus conservatifs (plus grands) en ajustant le niveau $\alpha' < \alpha$ définissant les valeurs critiques $t_{\alpha'/2;(n-J)}$ (Bonferroni $\alpha' = \alpha/(J(J-1)/2)$, Sidak). Dans le même esprit, la méthode de Scheffe, la plus conservative, projette l'ellipsoïde de confiance des moyennes des μ_i en intervalles de confiance des différences ou de toute combinaison linéaire de celles-ci (contrastes).

D'autres procédures définissent des intervalles studentisés fournissant des valeurs critiques spécifiques qui sont tabulées ou calculées par le logiciel. Certaines de ces méthodes ou certaines présentations graphiques des résultats sont uniquement adaptées au cas équirépété (Tukey) tandis que d'autres sont adaptées à des classes présentant des effectifs différents (GT2, Gabriel).

En résumé, pour comparer toutes les moyennes dans le cas équirépété, les méthodes de Tukey ou Scheffe sont utilisées, celle de Bonferroni convient encore au cas déséquilibré. Pour comparer les moyennes à celle d'une classe ou traitement témoin, la méthode de Bonferroni ($\alpha' = \alpha/(J(J-1)/2)$) est encore utilisée tandis que Dunnet remplace Tukey dans le cas équirépété.

2.4 Homogénéité de la variance

Une hypothèse importante du modèle induit par l'analyse de variance est l'homogénéité des variances de chaque groupe. Conjointement à l'estimation du modèle et en supposant la normalité, il peut être instructif de contrôler cette homogénéité par un test de l'hypothèse

$$H_0 : \sigma_1^2 = \cdots = \sigma_J^2.$$

Bartlett a proposé le test suivant. Posons

$$M = \sum_{j=1}^{J} (n_j - 1) \ln(s^2/s_j^2)$$

et

$$c = \frac{1}{3(J-1)} \left(\sum_{j=1}^{J} (\frac{1}{n_j - 1}) - 1/\sum_{j=1}^{J} (n_j - 1) \right).$$

Sous H_0 et pour de grands échantillons, la statistique $M/(c+1)$ suit un χ^2 à $(J-1)$ degrés de liberté. Dans les mêmes conditions, une approximation peut être fournie par la statistique

$$F = \frac{dM}{(J-1)(d/f - M)},$$

avec

$$f = (1-c) + 2/d \text{ et } d = (J+1)/c^2,$$

qui suit une loi de Fisher à $(J-1)$ et d degrés de liberté.

Néanmoins ce test n'est pas robuste à la violation de l'hypothèse de normalité. C'est pourquoi il lui est préféré la méthode de Levene qui considère les variables :

$$Z_{ij} = |y_{ij} - \bar{y}_{.j}|$$

sur lesquelles est calculée une analyse de variance. Malgré que les Z_{ij} ne soient ni indépendantes ni identiquement distribuées suivant une loi normale, la statistique de Fisher issue de l'ANOVA fournit un test raisonnable de l'homoscédasticité.

Le graphique représentant le nuage des résidus ou les boites à moustaches en fonction des niveaux du facteur complète très utilement le diagnostic. En cas d'hétéroscédasticité et comme en régression, une transformation de la variable à expliquer Y ($\sqrt{Y}, \ln(Y), 1/Y \ldots$) permet de limiter les dégâts.

2.5 Tests non paramétriques

Lorsque l'hypothèse de normalité n'est pas satisfaite et que la taille trop petite de l'échantillon ne permet pas de supposer des propriétés asymptotiques, une procédure non-paramétrique peut encore être mise en œuvre. Elles sont des alternatives plausibles au test de Fisher pour tester l'égalité des moyennes.

La procédure la plus utilisée est la construction du test de Kruskal-Wallis basée sur les rangs. Toutes les observations sont ordonnées selon les valeurs y_{ij} qui sont remplacées par leur rang r_{ij}, les ex æquo sont remplacés par leur rang moyen. On montre que la statistique de ce test, construite sur la somme des rangs à l'intérieur de chaque groupe, suit asymptotiquement une loi du χ^2 à $(J-1)$ degrés de liberté.

Une autre procédure, utilisant cette fois des rangs normalisés $(a_{ij} = r_{ij}/(n+1))$ conduit à une autre statistique utilisée dans le test de van der Waerden.

3 Modèle à deux facteurs

La considération de deux (ou plus) facteurs explicatifs, dans un modèle d'analyse de variance, engendre plusieurs complications. La première concerne la notion d'*interaction* entre variables explicatives. D'autres seront introduites dans la section suivante. Cette section décrit le cas de deux facteurs explicatifs croisés c'est-à-dire dont les niveaux d'un facteur ne sont pas conditionnés par ceux de l'autre. Les niveaux du premier facteur sont notés par un indice j variant de 1 à J, ceux du deuxième par un indice k variant de 1 à K.

Pour chaque combinaison, on observe un même nombre $n_{jk} = c > 1$ de répétitions ce qui nous place dans le cas particulier d'un plan **équilibré** ou **équirépété**. Ceci introduit des simplifications importantes dans les estimations des paramètres ainsi que dans la décomposition des variances. Le cas plus général est évoqué dans la section suivante.

3.1 Modèle complet

On peut commencer par écrire un modèle de variance à un facteur présentant $J \times K$ niveaux (j,k) :

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk} \text{ où } \left\{ \begin{array}{lcl} j & = & 1, \ldots, J\,; \\ k & = & 1, \ldots, K\,; \\ i & = & 1, \ldots, n_{jk} = c\,; \end{array} \right.$$

en supposant que les termes d'erreur ε_{ijk} sont mutuellement indépendants et de même loi. Chacun des paramètres μ_{jk} est estimé sans biais par la moyenne

$$\bar{y}_{.jk} = \frac{1}{c}\sum_{i=1}^{c} y_{ijk}.$$

Définissons également les moyennes suivantes :

$$\begin{aligned}
\bar{y}_{.j.} &= \frac{1}{K}\sum_{k=1}^{K} \bar{y}_{.jk}, \\
\bar{y}_{..k} &= \frac{1}{J}\sum_{j=1}^{J} \bar{y}_{.jk}, \\
\bar{y}_{...} &= \frac{1}{J}\sum_{j=1}^{J} \bar{y}_{.j.} = \frac{1}{K}\sum_{k=1}^{K} \bar{y}_{..k}.
\end{aligned}$$

qui n'ont de sens que dans le cas équirépété. La même convention du point en indice est également utilisée pour exprimer les moyennes des paramètres μ_{ijk}.

Les moyennes de chaque cellule sont alors décomposées en plusieurs termes afin de faire apparaître l'influence de chaque facteur :

Terme	*Paramètre*	*Estimation*
Moyenne générale	$\mu_{..}$	$\bar{y}_{...}$
Effet niveau j du 1er facteur	$\alpha_j = \mu_{j.} - \mu_{..}$	$\bar{y}_{.j.} - \bar{y}_{...}$
Effet niveau k du 2ème facteur	$\beta_k = \mu_{.k} - \mu_{..}$	$\bar{y}_{..k} - \bar{y}_{...}$
Effet de l'*interaction*	$\gamma_{jk} = \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..}$	$\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + y_{...}$

Avec les notations du tableau ci-dessus, on appelle $\mu_{..}$ l'effet général, $\mu_{j.}$ l'effet du niveau j du premier facteur, α_j l'effet différentiel du niveau j du premier facteur (même chose avec $\mu_{.k}$ et β_k pour le 2ème facteur), γ_{jk} l'effet d'interaction des niveaux j et k.

Un modèle d'analyse de variance à deux facteurs s'écrit alors :

$$y_{ijk} = \mu_{..} + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk} \text{ où } \left\{ \begin{array}{rcl} j &=& 1,\ldots,J\,; \\ k &=& 1,\ldots,K\,; \\ i &=& 1,\ldots,n_{jk}=c\,; \end{array} \right.$$

avec les contraintes

$$\sum_{j=1}^{J}\alpha_j = \sum_{k=1}^{K}\beta_k = 0\,;\ \forall k, \sum_{j=1}^{J}\gamma_{jk} = 0\,;\ \forall j, \sum_{k=1}^{K}\gamma_{jk} = 0$$

qui découlent de la définition des effets et assurent l'unicité de la solution.

3.2 Interaction

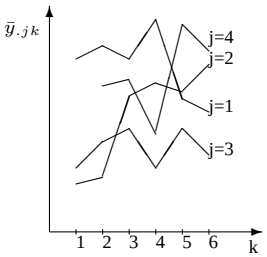
Lorsque les paramètres d'interaction γ_{jk} sont tous nuls, le modèle est dit *additif* ce qui correspond à une situation très particulière. Elle intervient lorsque

$$\bar{y}_{.jk} - \bar{y}_{..k} = \bar{y}_{.j.} - y_{...} \quad \forall j = 1,\ldots J\,;\ \forall k = 1,\ldots,K$$

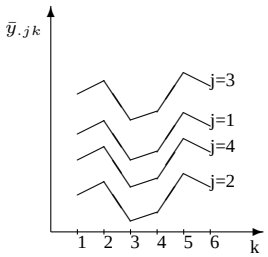
ce qui signifie que les écarts relatifs du premier facteur sont indépendants du niveau k du 2ème facteur (et vice versa).

Graphiquement, cela se traduit dans la figure 3.1 qui illustre les comportements des moyennes des cellules de modèles avec ou sans interaction (additif). Chaque ligne est appelée un *profil*, et la présence d'interactions se caractérise par le croisement de ces profils tandis que le parallélisme indique l'absence d'interactions. La question est évidemment de tester si des croisements observés sont jugés significatifs.

Attention, un manque de parallélisme peut aussi être dû à la présence d'une relation non-linéaire entre la variable Y et l'un des facteurs.



Avec interaction



Sans interaction

FIG. 3.1 – Moyennes de la variable Y pour chaque niveau d'un facteur en fonction des niveaux de l'autre facteur.

3.3 Modèles de régression

Comme dans le cas du modèle à un facteur, l'analyse d'un plan à deux facteurs se ramène à l'estimation et l'étude de modèles de régression sur variables indicatrices. En plus de celles des niveaux des deux facteurs $\{\mathbf{1}_1^1, \ldots, \mathbf{1}_J^1\}$, et $\{\mathbf{1}_1^2, \ldots, \mathbf{1}_K^2\}$, la prise en compte de l'interaction nécessite de considérer les indicatrices de chaque cellule ou traitement obtenues par produit des indicatrices des niveaux associés :

$$\mathbf{1}_{jk}^{1\times 2} = \mathbf{1}_j^1 \times \mathbf{1}_k^2 \,; j = 1 \ldots, J \,; k = 1, \ldots, K.$$

Le modèle s'écrit alors avec une autre paramétrisation :

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_{1,1}\mathbf{1}_1^1 + \cdots + \beta_{1,J}\mathbf{1}_J^1 + \beta_{2,1}\mathbf{1}_1^2 + \cdots + \beta_{2,K}\mathbf{1}_K^2 + \beta_{1\times 2,1}\mathbf{1}_1^{1\times 2} + \cdots + \beta_{1\times 2,JK}\mathbf{1}_{J,K}^{1\times 2} + \mathbf{u},$$

il comporte $1 + I + J + I \times J$ paramètres mais les colonnes de $\mathbf{X}$ sont soumises à de nombreuses combinaisons linéaires : une par paquet de $\mathbf{1}_j^1$ ou de $\mathbf{1}_k^2$ et une pour chaque paquet de $\mathbf{1}_{jk}^{1\times 2}$ à j ou k fixé. La matrice $\mathbf{X}'\mathbf{X}$ n'est pas inversible. Différentes approches sont proposées pour résoudre ce problème d'identifiabilité des paramètres.

- Supprimer une des indicatrices : en fonction de la base d'indicatrices choisie, différents modèles et donc différentes paramétrisations sont considérées.
- Ajouter une contrainte sur les paramètres afin de rendre unique la solution.
- Chercher une solution à partir d'un *inverse généralisé*[1] de la matrice $\mathbf{X}'\mathbf{X}$.

Dans le cas du modèle d'analyse de variance à *un* facteur, seule la première solution est couramment employée. Les autres, plus générales, le sont dans le cas de plusieurs facteurs et justifiées par des plannifications plus complexes ; différents inverses généralisés permettant de reconstruire les solutions avec contraintes ou par élimination d'une variable indicatrice. Les différents modèles considérés par les logiciels conduisent alors à des tests équivalents mais attention, la matrice $\mathbf{X}$ et le vecteur β prennent des significations différentes.

3.4 Stratégie de test

Une première décomposition de la variance associée au test général de nullité de tous les paramètres est proposée dans les logiciels mais celle-ci ne présente que peu d'intérêt. On considère ensuite les sommes de

[1] On dit que la matrice $\mathbf{A}^-$ est inverse généralisé de la matrice carrée $\mathbf{A}$ si elle vérifie : $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}^-$.

carrés spécifiques au cas équirépété :

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{...})^2 &&= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2 - cJK\bar{y}_{...}^2,\\
\text{SS1} &= cK\sum_{j=1}^{J}(\bar{y}_{.j.}-\bar{y}_{...})^2 &&= cK\sum_{j=1}^{J} y_{.j.}^2 - cJK\bar{y}_{...}^2,\\
\text{SS2} &= cJ\sum_{k=1}^{K}(\bar{y}_{..k}-\bar{y}_{...})^2 &&= cJ\sum_{k=1}^{K} y_{..k}^2 - cJK\bar{y}_{...}^2,\\
\text{SSI} &= c\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2 &&= c\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{.jk}^2 - cK\sum_{j=1}^{J} y_{.j.}^2 -\\
& && \quad -cJ\sum_{k=1}^{K} y_{..k}^2 + cJK\bar{y}_{...}^2,\\
\text{SSE} &= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{.jk})^2 &&= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2 - c\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{.jk}^2.
\end{aligned}$$

Dans ce cas, il est facile de montrer que tous les “doubles produits” des décompositions s’annulent (théorème de Pythagore) et

$$\text{SST} = \text{SS1} + \text{SS2} + \text{SSI} + \text{SSE}.$$

On parle alors de plans *orthogonaux* et les trois hypothèses suivantes (associées à des regroupements de contrastes) peuvent être considérées de façon indépendante :

$$\begin{aligned}
H_{03} &: \quad \gamma_{11} = \cdots = \gamma_{JK} = 0, && \text{pas d'effet d'interaction.}\\
H_{02} &: \quad \beta_1 = \cdots = \beta_K = 0, && \text{et } H_{03}\text{, pas d'effet du 2ème facteur}\\
H_{01} &: \quad \alpha_1 = \cdots = \alpha_J = 0, && \text{et } H_{03}\text{, pas d'effet du 1er facteur}
\end{aligned}$$

Elles sont évaluées dans la table ci-dessous qui présente l’unique décomposition de la variance dans le cas équirépété[2].

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	F
1er facteur	$J-1$	SS1	MS1=SS1/$(J-1)$	MS1/MSE
2ème facteur	$K-1$	SS2	MS2=SS2/$(K-1)$	MS2/MSE
Interaction	$(J-1)(K-1)$	SSI	MSI=$\frac{\text{SSI}}{(J-1)(K-1)}$	MSI/MSE
Erreur	$JK(c-1)$	SSE	MSE=SSE/$JK(c-1)$	
Total	$cJK-1$	SST		

Différentes stratégies de test sont suivies dans la littérature mais la plus couramment pratiquée consiste à comparer le modèle complet avec chacun des sous-modèles :

- Évaluer H_{03} de présence ou absence des termes d’interaction. Il existe des modèles intermédiaires de structuration de l’interaction mais le cas le plus simple du “tout ou rien” est retenu. Deux possibilités se présentent alors.
  i. Si l’interaction est significativement présente alors les deux facteurs sont influents ne serait-ce que par l’interaction. Il n’y a pas lieu de tester leur présence par H_{01} et H_{02}. Néanmoins il est d’usage de comparer les différentes statistiques F de test afin d’apprécier les rapports d’influence entre les effets principaux et l’interaction.

[2] Les options SS1,SS2, SS3, SS4 de SAS fournissent ainsi les mêmes résultats.

ii. Si l'interaction n'est pas significativement présente, il reste alors à tester l'effet de chaque facteur. Certains auteurs ré-estiment le modèle *additif* sans paramètre d'interaction (cf. remarque ci-dessous). Cela est déconseillé pour se protéger contre un manque possible de puissance du test de l'interaction. En effet, une faible interaction non décelée fausse l'estimation s^2 de σ_2. Il est donc préférable de conserver le modèle complet et de tester l'influence des facteurs par la nullité des α_j et β_j à partir des statistiques de la table ci-dessus.

Remarques

i. Si, compte tenu de connaissances *a priori* liées à un problème spécifique, l'interaction est éliminée du modèle, on est donc conduit à estimer un modèle additif plus simple (sans paramètres γ_{jk}). Dans ce cas, le nombre de paramètres à estimer et ainsi le nombre de degrés de liberté, la somme de carrés SSE et donc l'estimation $s^2 = MSE$ de σ^2 ne sont plus les valeurs fournies par la table d'analyse de variance ci-dessus. On distingue donc le cas d'un modèle *a priori* additif d'un modèle dans lequel l'hypothèse de nullité des interactions est acceptée.

ii. D'autres tests plus spécifiques sont construits en considérant des combinaisons linéaires des paramètres (contrastes) ou en effectuant des comparaisons multiples comme dans le cas à un facteur (Bonferroni, Tukey, Scheffe…).

iii. Les tests d'homogénéité des variances se traitent comme dans le cas du modèle à un facteur en considérant les JK combinaisons possibles.

4 Problèmes spécifiques

Certaines contraintes expérimentales peuvent induire des spécificités dans la planification et ainsi, par conséquence, dans le modèle d'analyse de variance associé. Un exposé détaillé des situations possibles sort du cadre de ce cours de 2ème cycle. Nous nous contenterons de citer ici certains problèmes courants en soulignant les difficultés occasionnées et quelques éléments de solution.

4.1 Facteur bloc

Les facteurs peuvent jouer des rôles différents. Certains sont contrôlés par l'expérimentateur qui sait en fixer précisément le niveau, d'autres, appelés *blocs*, sont des sources de variation propres aux procédés expérimentaux mais dont il faut tenir compte dans l'analyse car source d'hétérogénéité. L'exemple le plus typique concerne l'expérimentation agronomique en plein champ dans laquelle il est impossible de garantir l'homogénéité des conditions climatiques, hydrométriques ou encore de fertilité. Chaque champ ou bloc est donc découpé en parcelles "identiques" qui recevront chacune un traitement.

Dans d'autres situations, certaines mesures ne sont pas indépendantes, par exemple, lorsqu'elles sont réalisées sur les mêmes individus dans le cas de *mesures répétées*. Il est alors indispensable d'introduire un facteur bloc rendant compte de la structure particulière de l'expérimentation.

L'objectif est de séparer pour contrôler "au mieux" les sources de variation. Une "randomisation", ou tirage au sort, est réalisé à l'intérieur de chaque bloc afin de répartir "au hasard", dans l'espace, dans le temps, l'expérimentation des traitements ou combinaisons des autres facteurs.

4.2 Plan sans répétition

Si une seule expérience ou mesure est réalisée par cellule ou traitement, les composantes d'interaction et résiduelles sont confondues. Aucune hypothèse n'est testable dans le cadre général précédent. Il est néanmoins possible de se placer dans le cadre du modèle additif afin de tester l'influence de chaque facteur sous l'hypothèse implicite de non interaction.

4.3 Plans déséquilibrés, incomplets

Le cas de plans déséquilibrés, c'est-à-dire dans lesquels le nombre d'observations n'est pas le même dans chaque cellule ou pour chaque traitement, nécessite une attention particulière, surtout si, en plus, des cellules sont vides. Différents problèmes surgissent alors :

- les moyennes $\bar{y}_{.j.}$ ou $\bar{y}_{..k}$ définissant les estimateurs n'ont plus de sens,
- les "doubles produits" des décompositions des sommes de carrés ne se simplifient plus, il n'y a plus "orthogonalité",
- en conséquence, les hypothèses précédentes ou ensembles de contrastes ne peuvent plus être considérés de manière indépendante.

Néanmoins, l'approche générale par modèle linéaire des indicatrices reste valide. La solution obtenue par inverse généralisé :

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

n'est pas unique mais est utilisée pour construire des fonctions estimables des éléments de $\mathbf{b}$: $k'\mathbf{b}$ où k est un vecteur définissant un contraste. Plusieurs contrastes linéairement indépendants étant regroupés dans une matrice $\mathbf{K}$, l'hypothèse associée : $\mathbf{K}'\mathbf{b} = 0$ est alors testable en considérant la somme des carrés

$$SSK = (\mathbf{K}'\mathbf{b})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}]^{-1}(\mathbf{K}'\mathbf{b})$$

avec rang$(\mathbf{K})$ pour nombre de degrés de liberté.

Cette procédure "à la main" de construction des tests étant assez lourde, SAS propose l'étude d'hypothèses "standards" à travers quatre options. La première (SS1) fait intervenir l'ordre d'introduction des facteurs et est plus particulièrement adaptée aux modèles hiérarchisés, par exemple polynômiaux. La troisième (SS3) est conseillée dans les cas où les inégalités d'effectifs n'ont pas de signification particulière, ne sont pas dépendantes des niveaux des facteurs. Les deux autres options (SS2, SS4) ne sont guère utilisées, SS4, prévue pour les plans incomplets peut fournir des résultats étranges. En pratique standard, SS1 et SS3 sont comparées afin de s'assurer ou non de l'équirépartition puis les résultats de SS3 sont interprétés comme dans le cas équirépété.

4.4 Modèles à plus de deux facteurs

La prise en compte de plus de deux facteurs dans un modèle d'analyse de variance n'introduit pas de problème théorique fondamentalement nouveau. Seule la multiplication des indices et l'explosion combinatoire du nombre d'interactions à considérer compliquent la mise en œuvre pratique d'autant que beaucoup d'expérimentations sont nécessaires si la réalisation d'un plan complet est visée. Dans le cas contraire, tous les niveaux d'interaction ne sont pas testables et, comme dans le cas sans répétition, il faudra considérer des modèles moins ambitieux en supposant implicitement des hypothèses sur l'absence d'interactions d'ordres élevés. Si les facteurs sont très nombreux, il est courant de limiter chacun à 2 (ou 3 pour un modèle quadratique) niveaux et de ne considérer que certaines combinaisons deux à deux de facteurs. On parle alors de plans *fractionnaires*.

4.5 Facteurs hiérarchisés

Certains facteurs ou blocs peuvent par ailleurs être hiérarchisés ou emboîtés : les niveaux de certains facteurs sont conditionnés par d'autres facteurs. La composante d'interaction se confond alors avec la composante relative au facteur subordonné. Le modèle d'analyse de variance adapté à cette situation est dit *hiérarchisé*. Dans SAS, une syntaxe particulière permet de définir la structure.

5 Analyse de covariance

L'*analyse de covariance* se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Dans les cas les plus complexes, ont peut avoir plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynômiale. Le principe général est toujours d'estimer des modèles "*intra-groupes*" et de faire apparaître (tester) des effets différentiels "*inter-groupes*" des paramètres des régressions. Ainsi, dans le cas plus simple où seulement une variable parmi les explicatives est quantitative, nous sommes amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

5.1 Modèle

Le modèle est explicité dans le cas élémentaire où une variable quantitative Y est expliquée par une variable qualitative T à J niveaux et une variable quantitative, appelée encore covariable, X. Pour chaque niveau j de T, on observe n_j valeurs $x_{1j}, \ldots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de Y ; $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon.

En pratique, avant de lancer une procédure de modélisation et tests, une démarche exploratoire s'appuyant sur une représentation en couleur (une par modalité j de T) du nuage de points croisant Y et X et associant les droites de régression permet de se faire une idée sur les effets respectifs des variables : parallélisme des droites, étirement, imbrication des sous-nuages.

On suppose que les moyennes conditionnelles $E[Y|T]$, c'est-à-dire calculées à l'intérieur de chaque cellule, sont dans le sous-espace vectoriel engendré par les variables explicatives quantitatives, ici X. Ceci s'écrit :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \; ; \quad j = 1, \ldots, J; \quad i = 1, \cdots, n_j$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij}|i = 1, n_j; j = 1, J]'$ mis en colonne, $\mathbf{x}$ le vecteur $[x_{ij}|i = 1, n_j; j = 1, J]'$, $\mathbf{u} = [\varepsilon_{ij}|i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. On note encore $\mathbf{x}.\mathbf{1}_j$ le produit terme à terme des deux vecteurs, c'est-à-dire le vecteur contenant les observations de $\mathbf{X}$ sur les individus prenant le niveau j de T et des zéros ailleurs.

La résolution simultanée des J modèles de régression est alors obtenue en considérant globalement le modèle :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

dans lequel $\mathbf{X}$ est la matrice $n \times 2J$ constituée des blocs $[\mathbf{1}_j|\mathbf{x}.\mathbf{1}_j] \; ; j = 1, \ldots, J$. L'estimation de ce modèle global conduit, par bloc, à estimer les modèles de régression dans chacune des cellules.

Comme pour l'analyse de variance, les logiciels opèrent une reparamétrisation faisant apparaître des effets différentiels par rapport au dernier niveau (SAS/GLM, SAS/INSIGHT) ou par rapport à un effet moyen (Systat), afin d'obtenir directement les bonnes hypothèses dans les tests. Ainsi, dans le premier cas, on considère la matrice de même rang (sans la Jème indicatrice)

$$\mathbf{X} = [\mathbf{1}|\mathbf{x}|\mathbf{1}_1|\cdots|\mathbf{1}_{J-1}|\mathbf{x}.\mathbf{1}_1|\cdots|\mathbf{x}.\mathbf{1}_{J-1}]$$

associée aux modèles :

$$y_{ij} = \beta_{0J} + (\beta_{0j} - \beta_{0J}) + \beta_{1J} x_{ij} + (\beta_{1j} - \beta_{1J}) x_{ij} + \varepsilon_{ij} \; ; \quad j = 1, \ldots, J-1; i = 1, \ldots, n_j.$$

5.2 Tests

Différentes hypothèses sont alors testées en comparant le modèle complet

$$\begin{aligned} \mathbf{y} = \beta_{0J}\mathbf{1} \quad &+ \quad (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \\ &+ \quad (\beta_{11} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u} \end{aligned}$$

à chacun des modèles réduits :

$$\begin{aligned} (i) \qquad & \mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \mathbf{u} \\ (ii) \qquad & \mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01} - \beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1} - \beta_{0J})\mathbf{1}_{J-1} + \\ & + (\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u} \\ (iii) \qquad & \mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + (\beta_{1j} - \beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1} - \beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u} \end{aligned}$$

par un test de Fisher. Ceci revient à considérer les hypothèses suivantes :

- H_0^i : pas d'interaction, $\beta_{11} = \cdots = \beta_{1J}$, les droites partagent la même pente β_{1J},
- H_0^{ii} : β_{1J}=0,

- H_0^{iii} :$\beta_{01} = \cdots = \beta_{0J}$, les droites partagent la même constante à l'origine β_{0J}.

On commence donc par évaluer i), si le test n'est pas significatif, on regarde ii) qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si i) n'est pas significatif, on s'intéresse à iii) pour juger de l'effet du facteur T.

5.3 Cas général

Ce cadre théorique et les outils informatiques (SAS/GLM) permettent de considérer des modèles beaucoup plus complexes incluant plusieurs facteurs, plusieurs variables quantitatives, voire des polynômes de celles-ci, ainsi que les diverses interactions entre qualitatives et quantitatives. Le choix du "bon" modèle devient vite complexe d'autant que la stratégie dépend, comme pour la régression linéaire multiple, de l'objectif visé :

descriptif : des outils multidimensionnels descriptifs (ACP, AFD, AFCM...) s'avèrent souvent plus efficaces pour sélectionner, en première approche, un sous-ensemble de variables explicatives avant d'opérer une modélisation,

explicatif : de la prudence est requise d'autant que les hypothèses ne peuvent être évaluées de façon indépendante surtout si, en plus, des cellules sont déséquilibrées ou vides,

prédictif : la recherche d'un modèle efficace, donc parcimonieux, peut conduire à négliger des interactions ou effets principaux lorsqu'une faible amélioration du R^2 le justifie et même si le test correspondant apparaît comme significatif. L'utilisation du C_p est possible mais en général ce critère n'est pas calculé et d'utilisation délicate pour définir ce qu'est le "vrai" modèle de référence. En revanche, le PRESS donne des indications pertinentes.

6 Exemple

6.1 Les données

Les données, extraites de Jobson (1991), sont issues d'une étude marketing visant à étudier l'impact de différentes campagnes publicitaires sur les ventes de différents aliments. Un échantillon ou "panel" de familles a été constitué en tenant compte du lieu d'habitation ainsi que de la constitution de la famille. Chaque semaine, chacune de ces familles ont rempli un questionnaire décrivant les achats réalisés.

Nous nous limitons ici à l'étude de l'impact sur la consommation de lait de quatre campagnes diffusées sur des chaînes locales de télévision. Quatre villes, une par campagne publicitaire, ont été choisies dans cinq différentes régions géographiques. Les consommations en lait par chacune des six familles par ville alors été mesurées (en dollars) après deux mois de campagne.

Les données se présentent sous la forme d'un tableau à 6 variables : la région géographique, les 4 consommations pour chacune des villes ou campagnes publicitaires diffusées, la taille de la famille.

6.2 Analyse de variance à un facteur

Une première étude s'intéresse à l'effet du simple facteur "type de campagne publicitaire". On suppose implicitement que les familles ont été désignées aléatoirement indépendamment de l'appartenance géographique ou de leur taille. La procédure SAS/ANOVA est utilisée dans le programme suivant. Elle plus particulièrement adaptée aux situations équilibrées comme c'est le cas pour cet exemple. Le cas déséquilibré ne pose pas de problème majeur pour un modèle à un facteur mais pour deux facteurs ou plus, un message signale que les résultats sont fournis sous la responsabilité de l'utilisateur. Dans ce cas, la procédure plus générale SAS/GLM doit être utilisée.

Après une réorganisation des données permettant de construire une nouvelle variable décrivant le facteur "pub" ainsi que la variable unique consommation, le programme suivant a été exécuté :

```
title;
options pagesize=66 linesize=110 nonumber nodate;
proc anova data=sasuser.milkcc;
class pub;
```

```
model consom=pub;
means pub/bon scheffe tukey;
run;
```

SAS/ANOVA estime les paramètres du modèle d'analyse de variance à un facteur puis présente ensuite les résultats des tests de comparaison multiple demandés en option. Cette procédure signale explicitement que des problèmes peuvent apparaître si certains tests, spécifiques au cas équirépété, sont utilisés hors de leur contexte. Différentes options de présentation des résultats sont proposées : tests avec niveau paramétrable (5% par défaut) de significativité, intervalles de confiance des différences ou des moyennes.

Dans cet exemple, une des trois procédures de tests utilisée ne conclut pas aux mêmes résultats. Les tests de Scheffe acceptent tous l'hypothèse H_0 d'égalité des différentes moyennes. on retrouve ainsi le caractère conservatif de cette procédure.

La procédure SAS/NPAR1WAY a ensuite été exécutée pour obtenir les résultats des test non-paramétriques.

```
proc npar1way data=sasuser.milkcc;
class pub;
var consom;
run;
```

Les résultats sont encore "mitigés".

```
Source            DF   Sum of Squares     Mean Square    F Value   Pr > F
                 (1)
Model              3   4585.68048667(2) 1528.56016222(5) 3.16(7)   0.0275(8)
Error            116  56187.44398000(3)  484.37451707(6)
Corrected Total  119  60773.12446667(4)

  R-Square            C.V.           Root MSE            CONSOM Mean
  0.075456(12)     54.05283(11)    22.00851011(9)       40.71666667(10)
```

(1)	degrés de liberté pour le calcul des moyennes et la sélection de la loi de Fisher du test global
(2)	SSB
(3)	SSW
(4)	SST=SSW+SSB
(5)	SSB/DF
(6)	s^2 =MSE=SSW/DF est l'estimation de σ_u^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10)
(12)	Coefficient de détermination R^2

```
Tukey's Studentized Range (HSD)
Alpha= 0.05  df= 116  MSE= 484.3745
Minimum Significant Difference= 14.813
Means with the same letter are not significantly different.

    Tukey Grouping              Mean      N  PUB

                A            51.030     30  4
                A
          B     A            39.647     30  2
          B     A
          B     A            37.239     30  1
          B
          B                  34.951     30  3
```

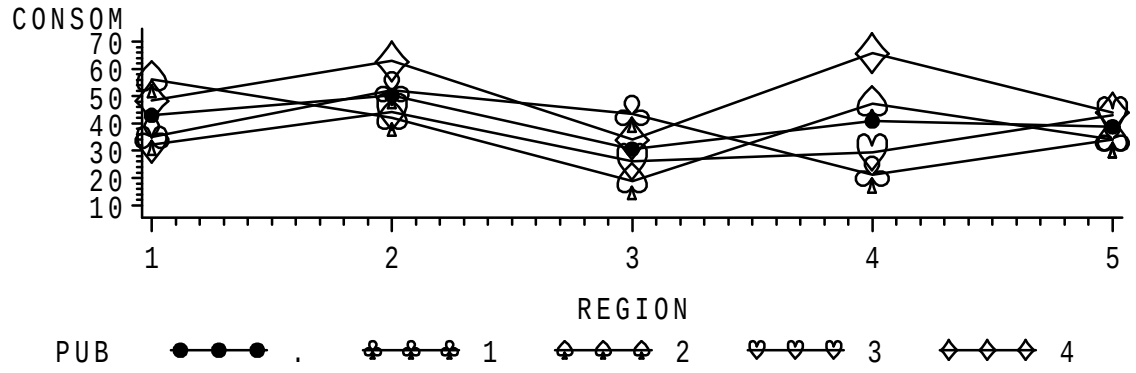


FIG. 3.2 – Profil moyen et profils de la consommation moyenne de chaque région en fonction du type de campagne.

Test non-paramétrique

```
Wilcoxon Scores (Rank Sums) for Variable CONSOM
Classified by Variable PUB
                    Sum of      Expected      Std Dev         Mean
PUB         N        Scores     Under H0     Under H0        Score
1          30    1675.00000        1815.0    164.999427    55.8333333
2          30    1781.50000        1815.0    164.999427    59.3833333
3          30    1562.50000        1815.0    164.999427    52.0833333
4          30    2241.00000        1815.0    164.999427    74.7000000

Kruskal-Wallis Test (Chi-Square Approximation)
CHISQ =  7.3266    DF =  3    Prob > CHISQ = 0.0622
```

6.3 Modèle à deux facteurs

Une étude graphique préalable des interactions est toujours instructive :

```
proc means data=sasuser.milkcc mean stderr;
class pub region;
var consom;
output out=cellmoy mean=moycons;
run;
symbol i=join v=dot cv=black ;
symbol2 i=join v=% cv=black h=2;
symbol3 i=join v='"' cv=black h=2;
symbol4 i=join v=# cv=black h=2;
symbol5 i=join v=$ cv=black h=2;%$
proc gplot data=cellmoy;
plot moycons*region=pub;
run;
goptions reset=all; quit;
```

Nous sommes dans le cas équirépété, la procédure SAS/ANOVA reste valide mais SAS/GLM, plus générale, est utilisée et fournit dans ce cas les mêmes résultats. Cette procédure adaptée aux situations complexes fournit également d'autres options (contrastes, estimation des paramètres…).

```
title;
options pagesize=66 linesize=110 nonumber nodate;
proc glm data=sasuser.milkcc;
class pub region;
```

```
model consom= pub region pub*region;

run;
```

```
General Linear Models Procedure
 (0)
Source            DF     Sum of Squares     Mean Square   F Value   Pr > F
Model             19     18391.10933333    967.95312281   2.28      0.0045
Error             100    42382.01513333    423.82015133
Corrected Total   119    60773.12446667
    R-Square        C.V.        Root MSE         CONSOM Mean
    0.302619      50.56134    20.58689271        40.71666667

Source            DF     Type III SS        Mean Square   F Value   Pr > F
                  (1)                           (5)         (6)      (7)
PUB                3     4585.68048667(2)  1528.56016222   3.61     0.0160
REGION             4     4867.51141667(3)  1216.87785417   2.87     0.0268
PUB*REGION        12     8937.91743000(4)  744.82645250    1.76     0.0658
```

(0)	Tableau associé au test global de nullité de tous les paramètres.
(1)	Degrés de liberté pour le calcul des moyennes et sélection de la loi de Fisher.
(2)	SS1
(3)	SS2
(4)	SSI
(5)	SS1,2,I/DF
(6)	Statistique F pour chacun des tests
(7)	$P(f_{p;n-p-1} > F)$; H_i est rejetée au niveau α si $P < \alpha$

6.4 Analyse de covariance

La variable “taille” est quantitative. On s’intéresse à différents modèles de régression visant à expliquer la consommation en fonction de la taille de la famille conditionnellement au type de campagne publicitaire.

```
proc glm data=sasuser.milk;
class pub;
model consom=pub taille pub*taille;
run;
```

Les résultats ci-dessous conduiraient à conclure à une forte influence de la taille mais à l’absence d’influence du type de campagne. Les droites de régression ne semblent pas significativement différentes.

```
Source         DF   Type III SS   Mean Square   F Value     Pr > F
PUB            3      227.1807       75.7269      0.57     0.6377 (1)
TAILLE         1    40926.0157    40926.0157    306.57     0.0001 (2)
TAILLE*PUB     3      309.8451      103.2817      0.77     0.5111 (3)
```

(1)	Test de la significativité des différences des termes constants.
(2)	Test de l’influence du facteur quantitatif.
(3)	Test de la significativité des différences des pentes (interaction).

Néanmoins, compte tenu des résultats précédents (analyse de variance), le même calcul est effectué pour chaque région :

```
proc glm data=sasuser.milk;
by region;
class pub;
```

```
model consom=pub taille pub*taille;
run;
```

Région	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	PUB	3	72.02974	24.00991	4.62	0.0164
1	TAILLE	1	7178.32142	7178.32142	1380.25	0.0001
	TAILLE*PUB	3	217.37048	72.45683	13.93	0.0001
	PUB	3	231.73422	77.24474	30.36	0.0001
2	TAILLE	1	8655.25201	8655.25201	3402.34	0.0001
	TAILLE*PUB	3	50.15069	16.71690	6.57	0.0042
	PUB	3	79.54688	26.51563	6.01	0.0061
3	TAILLE	1	6993.30160	6993.30160	1585.35	0.0001
	TAILLE*PUB	3	173.19305	57.73102	13.09	0.0001
	PUB	3	415.66664	138.55555	15.23	0.0001
4	TAILLE	1	9743.37830	9743.37830	1071.32	0.0001
	TAILLE*PUB	3	361.39556	120.46519	13.25	0.0001
	PUB	3	15.35494	5.11831	0.79	0.5168
5	TAILLE	1	8513.28516	8513.28516	1314.71	0.0001
	TAILLE*PUB	3	52.75119	17.58373	2.72	0.0793

Il apparaît alors qu'à l'intérieur de chaque région (sauf région 5), les campagnes de publicité ont un effet tant sur la constante que sur la pente.

Ceci incite donc à se méfier des *interactions* et encourage à toujours conserver le facteur bloc dans une analyse. Une approche complète, considérant *a priori* toutes les variables (3 facteurs), est ici nécessaire (cf. TP).

7 Exercices

Exo 1

On se place dans le cadre du modèle d'analyse de variance à un facteur :

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, n_j; j = 1, J.$$

On se propose de comparer la régression de la variable Y sur deux systèmes d'indicatrices $\mathbf{1}_j$ des modalités engendrant le même espace. On rappelle les formules d'inversion par bloc d'une matrice carrée $\mathbf{A}^{-1} = \mathbf{B}$:

i. $\mathbf{B}_{11} = [\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{B}_{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}$

ii. $\mathbf{B}_{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}[\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}]^{-1} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{B}_{22}$

iii. $\mathbf{B}_{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}[\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11}$

iv. $\mathbf{B}_{22} = [\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}]^{-1} = \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$

Pour chacun des deux modèles de régression ci-dessous :

$$\mathbf{y} = \beta_1\mathbf{1}_1 + \cdots + \beta_J\mathbf{1}_J + \mathbf{u} \tag{3.1}$$

$$\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{1}_1 + \cdots + \beta_{J-1}\mathbf{1}_{J-1} + \mathbf{u} \tag{3.2}$$

exprimer les matrices $\mathbf{X}, \mathbf{X}'\mathbf{X}, (\mathbf{X}'\mathbf{X})^{-1}$ associées ainsi que le vecteur $\mathbf{b}$ des estimations des paramètres.

Exo 2

Considérons le modèle équilibré d'analyse de variance à deux facteurs :

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk} \text{ où } \begin{cases} j = 1, \ldots, J\,; \\ k = 1, \ldots, K\,; \\ i = 1, \ldots, n_{jk} = c\,; \end{cases}$$

Les ε_{ijk} sont supposés mutuellement indépendants et de même distribution $\mathcal{N}(0, \sigma^2)$.

i. Montrer la décomposition des sommes de carrés : SST=SS1+SS2+SSI+SSE.

ii. Exprimer $E[\bar{y}^2_{..k}], E[\bar{y}^2_{.j.}], E[\bar{y}^2_{.jk}], E[\bar{y}^2_{...}]$ en fonction de σ^2, J, K et des paramètres $\mu, \mu_{j.}, \mu_{.k}, \mu_{jk}$ (calculer d'abord les moyennes puis les variances).

iii. En déduire que

$$\begin{aligned} E[SS1] &= \sigma^2(J-1) + cK\sum_{j=1}^{J}(\mu_{j.}-\mu)^2 \\ E[SS2] &= \sigma^2(K-1) + cJ\sum_{k=1}^{K}(\mu_{.k}-\mu)^2 \\ E[SSI] &= \sigma^2(J-1)(K-1) + c\sum_{j=1}^{J}\sum_{k=1}^{K}(\mu_{jk}-\mu_{j.}-\mu_{.k}+\mu)^2 \\ E[SSE] &= \sigma^2 JK(c-1) \end{aligned}$$

Exo 3

Un agronome a mesuré le rendement d'une culture de haricots en fonction de deux caractères : la variété de l'espèce (5 niveaux), et un traitement (3 niveaux). Il obtient 15 observations rangées dans le tableau ci-dessous.

	A	B	C	D	E
T1	17.5	20.0	18.0	17.0	16.5
T2	15.1	16.0	13.0	12.0	14.5
T3	10.0	13.0	10.0	11.0	12.0

Il considère comme modèle de référence un modèle gaussien.

i. Qu'obtiendrait-il en ajustant un modèle à 2 facteurs avec interactions sur ces données (valeurs estimées, estimation de σ, résidus) ? Est-il possible de tester la présence d'interaction ?

ii. Il veut estimer le modèle à deux facteurs sans interaction avec SAS. Quelle procédure doit-il utiliser, `anova` ou `glm` ?

iii. La procédure `glm` lui fournit les résultats suivant :

```
                        General Linear Models Procedure
Dependent Variable: RENDMNT
                                    Sum of           Mean
Source                   DF        Squares         Square   F Value     Pr > F
Model                     6     125.74400000     20.95733333     20.63     0.0002
Error                     8       8.12533333      1.01566667
Corrected Total          14     133.86933333

                   R-Square             C.V.        Root MSE          RENDMNT Mean
                   0.939304         7.011616       1.0078029            14.373333
Source                   DF      Type III SS     Mean Square   F Value     Pr > F
TRAIT                     2     109.38133333     54.69066667     53.85     0.0001
VARIET                    4      16.36266667      4.09066667      4.03     0.0445
```

Quelles onclusions tirer sur l'effet des facteurs ?

iv. Le test de Bonferroni de comparaison multiple sur les variétés puis sur les traitements donnent les résultats ci-dessous :

```
Bon Grouping              Mean      N  TRAIT
           A           17.8000      5  T1
           B           14.1200      5  T2
           C           11.2000      5  T3

Bon Grouping              Mean      N  VARIET
           A          16.3333      3  B
           A
```

```
A          14.3333      3  E
A
A          14.2000      3  A
A
A          13.6667      3  C
A
A          13.3333      3  D
```

Que peut-il conclure ? Est-ce cohérent avec les résultats précédents ?

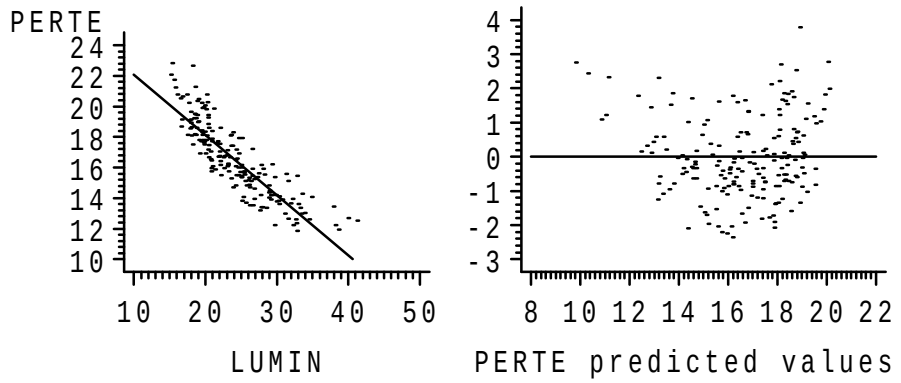
Exo4

El Ringo achète du café vert dans le monde entier avant de le torréfier puis de le redistribuer. Son problème est de prévoir la *perte de poids* due à la torréfaction. Cette perte, qui peut atteindre 20%, conditionne directement sa marge bénéficiaire, elle doit être estimée le plus précisément possible au moment de l'achat afin de pouvoir négocier le prix au plus juste. Son nez, légendaire lors de la torréfaction, est inneficace sur du café vert. El Ringo fait l'acquisition d'un chromatographe qui peut lui fournir rapidement 5 indicateurs numériques à partir d'un échantillon. Il réalise alors 189 expériences sur des échantillons de diverses provenances et construit un tableau contenant pour chaque échantillon les mesures chromatographiques sur le café vert (`lumin, xa, xb xy xgn`) et la `perte` de poids après torréfaction. L'objectif est de construire un bon modèle de prédiction.

i. Grâce à son tableur il tente d'expliquer la `perte` par la variable `lumin` et obtient les résultats ci-dessous. Critiquez ceux-ci.

```
modèle perte=lumin
   Model            1    837.16554    837.16554      581.768       0.0001
   Error          187    269.09343      1.43900
   C Total        188   1106.25897

       Root MSE       1.19958     R-square       0.7568
       Dep Mean      16.54958     Adj R-sq       0.7555
       C.V.           7.24843
```

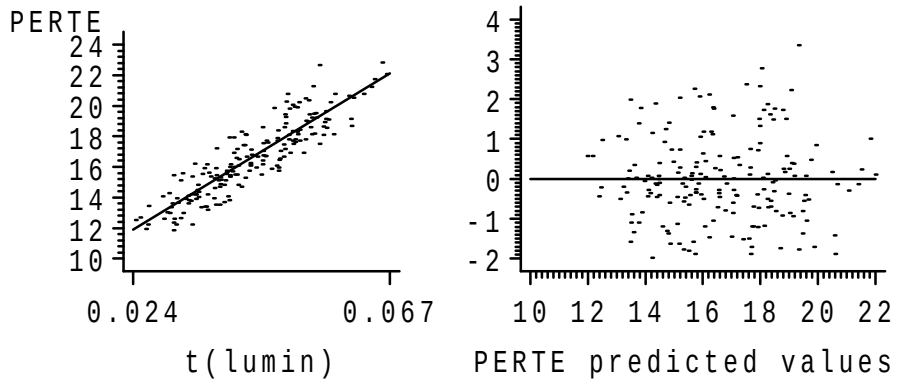


Régression de la `perte` *en fonction de* `lumin` *et graphe des résidus*

ii. Son voisin physicien lui conseille une autre analyse, utilisant une variable `Tlumin`, qui fournit les résultats suivants. Quelle est cette variable `Tlumin` ? Que devient le modèle ? Ce choix est-il fructueux ?

```
modèle perte=Tlumin
   Model            1    893.41575    893.41575      784.938       0.0001
   Error          187    212.84322      1.13820
   C Total        188   1106.25897

       Root MSE       1.06686     R-square       0.8076
       Dep Mean      16.54958     Adj R-sq       0.8066
       C.V.           6.44647
```



Régression de la `perte` *en fonction de* `t(lumin)` *et graphe des résidus*

iii. Le biologiste de l'entreprise recommande, compte tenu des enjeux financiers énormes, d'investir dans le logiciel TASS. Il suggère d'introduire toutes les variables initiales et certaines transformées par la même fonction T, dans un modèle de régression multiple et de ne retenir que le "meilleur". Que contient le tableau ci-dessous ? Quel modèle conseilleriez vous ?

```
The TASS System
modele perte=Tlumin lumin Txgn xgn Txy xy xa xb
         R-square      Adj     C(p)      BIC Variables in Model
    In                 Rsq
       1 0.876661 0.876001  63.9928 -56.6606 XA
       1 0.820516 0.819556    177.3  12.9508 XB
  ----------------------------------------------
       2 0.892518 0.891362  33.9808 -80.3970 XA XGN
       2 0.889934 0.888750  39.1974 -76.0409 XA XY
  ------------------------------------------------
       3 0.902013 0.900424  16.8117 -95.4315 LUMIN XA XB
       3 0.899415 0.897784  22.0572 -90.6881 XA XGN TXGN
  --------------------------------------------------
       4 0.905612 0.903560  11.5466   -100.2 LUMIN XA XB TXGN
       4 0.904280 0.902200  14.2352 -97.7154 LUMIN TLUMIN XA XB
  ----------------------------------------------------
       5 0.910376 0.907927   3.9302   -107.4 LUMIN XA XB XY TXGN
       5 0.907249 0.904715  10.2426   -101.3 LUMIN XA XB XGN TXGN
  ------------------------------------------------------
       6 0.910557 0.907608   5.5647   -105.7 LUMIN XA XB XY TXY TXGN
       6 0.910462 0.907510   5.7560   -105.5 LUMIN XA XB XY XGN TXGN
  --------------------------------------------------------
       7 0.910732 0.907279   7.2117   -103.9 LUMIN TLUMIN XA XB XY TXY TXGN
       7 0.910557 0.907098   7.5637   -103.6 LUMIN XA XB XY TXY XGN TXGN
  ----------------------------------------------------------
       8 0.910837 0.906874   9.0000   -102.0 LUMIN TLUMIN XA XB XY TXY XGN TXGN
```

iv. Un collègue du service financier remarque que le modèle ne tient pas compte de l'origine (Arabie, Afrique, Amérique…) du café vert alors que celle-ci est connue. Cette origine est codée de 1 à 7 dans une variable nommée `cafe`. Quelle méthode, quelle stratégie proposeriez vous afin de rechercher un éventuel meilleur modèle de prévision ? (5 lignes max).

v. La procédure de sélection adoptée passe par les étapes ci-dessous dont chacune est résumée par un tableau de décomposition de la variance. Pour chacune de ces trois étapes, indiquer le modèle a essayer dans l'étape suivante.

```
modele perte= cafe Tlumin lumin Txgn xgn Txy xy xa xb
                          Type III Tests
Source                DF  Sum of Squares  Mean Square      F Stat    Prob > F
CAFE                   5          1.8525       0.3705      1.2789      0.2771
LUMIN                  1          0.0284       0.0284      0.0981      0.7546
XA                     1          0.0070       0.0070      0.0242      0.8765
XB                     1          0.1649       0.1649      0.5692      0.4520
XY                     1          0.1286       0.1286      0.4439      0.5065
XGN                    1          0.1380       0.1380      0.4764      0.4913
TLUMIN                 1          0.1880       0.1880      0.6489      0.4220
TXY                    1          0.1845       0.1845      0.6367      0.4264
TXGN                   1          0.0071       0.0071      0.0245      0.8758
LUMIN*CAFE             6          2.8909       0.4818      1.6631      0.1354
```

XA*CAFE	6	2.8391	0.4732	1.6333	0.1432
XB*CAFE	6	4.1561	0.6927	2.3910	0.0320
XY*CAFE	6	3.6205	0.6034	2.0829	0.0597
XGN*CAFE	6	2.4229	0.4038	1.3939	0.2221
TLUMIN*CAFE	4	1.2644	0.3161	1.0911	0.3639
TXY*CAFE	5	2.1247	0.4249	1.4668	0.2053
TXGN*CAFE	6	2.4673	0.4112	1.4194	0.2122

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
CAFE	6	1.3293	0.2216	0.7696	0.5951
LUMIN	1	0.0069	0.0069	0.0240	0.8772
XA	1	0.0016	0.0016	0.0056	0.9402
XB	1	0.2542	0.2542	0.8830	0.3490
XY	1	0.2079	0.2079	0.7220	0.3970
XGN	1	0.1889	0.1889	0.6563	0.4193
TLUMIN	1	2.7040	2.7040	9.3920	0.0026
TXY	1	0.7597	0.7597	2.6388	0.1066
TXGN	1	0.0801	0.0801	0.2781	0.5988
XA*CAFE	6	4.9587	0.8265	2.8706	0.0115
XB*CAFE	6	5.5532	0.9255	3.2147	0.0055
XY*CAFE	6	3.6637	0.6106	2.1209	0.0547
XGN*CAFE	6	3.1080	0.5180	1.7992	0.1036
TXY*CAFE	6	3.0950	0.5158	1.7917	0.1051
TXGN*CAFE	6	3.2415	0.5402	1.8765	0.0890

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
XA	1	3.1569	3.1569	9.7080	0.0022
XY	1	7.1315	7.1315	21.9306	0.0001
TLUMIN	1	8.1339	8.1339	25.0131	0.0001
TXY	1	1.1789	1.1789	3.6254	0.0586
TXGN	1	1.0512	1.0512	3.2327	0.0740
XA*CAFE	6	3.9667	0.6611	2.0331	0.0640
TXGN*CAFE	6	4.9245	0.8208	2.5239	0.0231

Type III Tests

Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
CAFE	6	57.6901	9.6150	27.18	0.0001
TLUMIN	1	36.8761	36.8761	104.26	0.0001
XA	1	10.9798	10.9798	31.04	0.0001
XY	1	7.1634	7.1634	20.25	0.0001
TXGN	1	5.0804	5.0804	14.36	0.0002

vi. Commenter la structure du modèle obtenu, discuter l'effet de la variable `cafe` (5 lignes max).

vii. Le dernier modèle conduit aux estimations ci-dessous. Pour les valeurs observées suivantes : `café=2 lumin=38.52 xa=11.57 xb=29.22 xy=10.38 xgn=8.05`, calculer la prévision de perte.

Parameter Estimates

Variable	CAFE	DF	Estimate	Std Error	T Stat
INTERCEPT		1	-29.2777	6.0301	-4.8552
CAFE	1	1	-3.7673	0.4022	-9.3676
	2	1	0.5179	0.1951	2.6547
	3	1	-2.9310	0.3302	-8.8766
	4	1	-2.9151	0.3253	-8.9604
	5	1	-1.6011	0.2474	-6.4721
	6	1	-0.8837	0.2423	-3.6473
	7	0	0.0000	.	.
TLUMIN		1	783.8576	76.7671	10.2109
XA		1	1.5281	0.2743	5.5717
XY		1	0.4947	0.1099	4.5004
TXGN		1	-11.2462	2.9673	-3.7900

viii. La procédure `TMLG` de TASS ne fournit pas, comme dans la question 3, le C_p de Mallows. En revanche elle fournit le "PRESS". Commenter les résultats ci-dessous (5 lignes).

Variables	R^2	PRESS
lumin xa xb xy txgn	0.910376	105.2029
cafe lumin xa xb xy txgn	0.919257	99.7226
cafe tlumin xa xb txgn	0.942234	72.4155
cafe tlumin xa xy txgn	0.943091	71.2120
cafe tlumin xa xb xy txgn	0.943364	71.5279

Chapitre 4

Modèles de dénombrement

Les modèles décrits dans ce chapitre s'intéressent plus particulièrement à la description ou l'explication d'observations constitués d'effectifs comme, par exemple, le nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais ou encore le nombre d'individus qui prennent une combinaison donnée de modalités de variables qualitatives ou niveaux de facteurs.

Contrairement aux modèles des chapitres précédents basés sur l'hypothèse de normalité des observations, les lois concernées sont maintenant discrètes et associées à des dénombrements : loi de Poisson, binomiale, multinomiale. Néanmoins, tous les modèles considérés dans ce cours appartiennent à la famille des *modèles linéaires généralisés*. Dans ce chapitre, nous définissons le contexte pratique de la *régression logistique* et du *modèle log-linéaire* tandis que les aspects communs à ces deux techniques, (estimation, tests, diagnostic) et dont la stratégie de mise en œuvre est similaire au cas gaussien, sont détaillés dans l'introduction au modèle linéaire généralisé présentée dans le chapitre suivant. Une première section définit quelques notions relatives à l'étude de la liaison entre variables qualitatives. Elles sont couramment utilisées dans l'interprétation des modèles de ce chapitre.

1 Odds et odds ratio

Une variable

Soit Y une variable qualitative à J modalités. On désigne la chance ou *l'odds*[1] de voir se réaliser la $j$ème modalité plutôt que la $k$ème par le rapport

$$\Omega_{jk} = \frac{\pi_j}{\pi_k}$$

où π_j est la probabilité d'apparition de la $j$ème modalité. Cette quantité est estimée par le rapport n_j/n_k des effectifs observés sur un échantillon. Lorsque la variable est binaire et suit une loi de Bernouilli de paramètre π, l'odds est le rapport $\pi/(1-\pi)$ qui exprime une cote ou chance de gain.

Table de contingence

On considère maintenant une table de contingence 2×2 croisant deux variables qualitatives binaires X^1 et X^2. les paramètres de la loi conjointe se mettent dans une matrice :

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

où $\pi_{ij} = P[\{X^1 = i\} \text{ et } \{X^2 = j\}]$ est la probabilité d'occurence de chaque combinaison.

- Dans la ligne 1, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}}.$$

[1] Il n'existe pas, même en Québécois, de traduction consensuelle de "odds".

- Dans la ligne 2, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_2 = \frac{\pi_{21}}{\pi_{22}}.$$

On appelle odds ratio le rapport

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Ce rapport prend la valeur 1 si les variables sont indépendantes, il est supérieur à 1 si les sujets de la ligne 1 ont plus de chances de prendre la première colonne que les sujets de la ligne 2 et inférieur à 1 sinon.

L'odds ratio est également défini pour deux lignes (a, b) et deux colonnes (c, d) quelconques d'une table de contingence croisant deux variables à J et K modalités. L'odds ratio est le rapport

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}} \quad \text{estimé par l'odds ratio empirique} \quad \widehat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}.$$

2 Régression logistique

2.1 Type de données

Cette section décrit la modélisation d'une variable qualitative Z à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de maladie, panne d'un équipement, faillite d'une entreprise, bon ou mauvais client.... Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\beta$ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités

$$\pi = P(Z = 1) \quad \text{ou} \quad 1 - \pi = P(Z = 0),$$

ou plutôt une transformation de celles-ci, par l'observation conjointe des variables explicatives. L'idée est en effet de faire intervenir une fonction réelle monotone g opérant de $[0, 1]$ dans $\mathbb{R}$ et donc de chercher un modèle linéaire de la forme :

$$g(\pi_i) = \mathbf{x}_i'\beta.$$

Il existe de nombreuses fonctions, dont le graphe présente une forme sigmoïdale et qui sont candidates pour remplir ce rôle, trois sont pratiquement disponibles dans les logiciels :

probit : g est alors la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.

log-log avec g définie par

$$g(\pi) = \ln[-\ln(1 - \pi)]$$

mais cette fonction est dissymétrique.

logit est définie par

$$g(\pi) = \textit{logit}(\pi) = \ln\frac{\pi}{1 - \pi} \quad \text{avec} \quad g^{-1}(x) = \frac{e^x}{1 + e^x}.$$

Plusieurs raisons, tant théoriques que pratiques, font préférer cette dernière solution. Le rapport $\pi/(1 - \pi)$, qui exprime une "cote", est l'*odds* et la *régression logistique* s'interprète donc comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio" c'est-à-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.

Cette section se limite à la description de l'usage élémentaire de la régression logistique. Des compléments concernant l'explication d'une variable qualitative ordinale (plusieurs modalités), l'intervention de variables explicatives avec effet aléatoire, l'utilisation de mesures répétées donc dépendantes, sont à rechercher dans la bibliographie.

2.2 Modèle binomial

On considère, pour $i = 1, \ldots, I$, différentes valeurs *fixées* $x_i^1, \ldots, x_i^q$ des variables explicatives $X^1, \ldots, X^q$. Ces dernières pouvant être des variables quantitatives ou encore des variables qualitatives, c'est-à-dire des facteurs issus d'une planification expérimentale.

Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations ($n = \sum_{i=1}^{I} n_i$) de la variable Z qui se mettent sous la forme $y_1/n_1, \ldots, y_I/n_I$ où y_i désigne le nombre de "succès" observés lors des n_i essais. On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, la probabilité π_i de succès est constante. Alors, la variable Y_i sachant n_i et d'espérance $E(Y_i) = n_i\pi_i$ suit une loi *binomiale* $\mathcal{B}(n_i, \pi_i)$ dont la fonction de densité s'écrit :

$$P(Y = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{(n_i - y_i)}.$$

On suppose que le vecteur des fonctions *logit* des probabilités π_i appartient au sous-espace vect$\{X^1, \ldots, X^q\}$ engendré par les variables explicatives :

$$\textit{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, I$$

ce qui s'écrit encore

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} \quad i = 1, \ldots, I.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.

L'optimisation fournit une estimation $\mathbf{b}$ de β, il est alors facile d'en déduire les estimations ou prévisions des probabilités π_i :

$$\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1 + e^{\mathbf{x}_i'\mathbf{b}}}$$

et ainsi celles des effectifs

$$\widehat{y}_i = n_i\widehat{\pi}_i.$$

Remarques

i. La matrice $\mathbf{X}$ issue de la planification expérimentale est construite avec les mêmes règles que celles utilisées dans le cadre de l'analyse de covariance mixant variables explicatives quantitatives et qualitatives. Ainsi, les logiciels gèrent avec plus ou moins de clarté le choix des variables indicatrices et donc des paramètres estimables ou contrastes associés.

ii. La situation décrite précédemment correspond à l'observation de données *groupées*. Dans de nombreuses situations concrètes et souvent dès qu'il y a des variables explicatives quantitatives, les observations $\mathbf{x}_i$ sont toutes distinctes. Ceci revient donc à fixer $n_i = 1; i = 1, \ldots, I$ dans les expressions précédentes et la loi de Bernouilli remplace la loi binomiale. Certaines méthodes ne sont alors plus applicables et les comportements asymptotiques des distributions des statistiques de test ne sont plus valides, le nombre de paramètres tendant vers l'infini.

3 Modèle log-linéaire

3.1 Types de données

Les données se présentent généralement sous la forme d'une table de contingence obtenue par le croisement de plusieurs variables qualitatives et dont chaque cellule contient un effectif ou une fréquence à modéliser. Nous nous limiterons à l'étude d'une table élémentaire en laissant de côté des structures plus complexes, par exemple lorsque des zéros structurels, des indépendances conditionnelles, des propriétés de

symétrie ou quasi-symétrie, une table creuse, sont à prendre en compte. D'autre part, sous sa forme la plus générale, le modèle peut intégrer également des variables quantitatives.

Ce type de situation se retrouve en analyse des correspondances simple ou multiple mais ici, l'objectif est d'expliquer ou de modéliser les effectifs en fonction des modalités prises par les variables qualitatives. L'objectif final pouvant être *explicatif* : tester une structure de dépendance particulière, ou *prédictif* avec choix d'un modèle parcimonieux.

3.2 Distributions

On considère la table de contingence complète constituée à partir de l'observation des variables qualitatives $X^1, X^2, \ldots, X^p$ sur un échantillon de n individus. Les effectifs $\{y_{jk\ldots l}; j = 1, J; k = 1, K; \ldots; l = 1, L\}$ de chaque cellule sont rangés dans un vecteur $\mathbf{y}$ à $I(I = J \times K \times \cdots \times L)$ composantes. Différentes hypothèses sur les distributions sont considérées en fonction du contexte expérimental.

Poisson

Le modèle le plus simple consiste à supposer que les variables observées Y_i suivent des lois de Poisson indépendantes de paramètre $\mu_i = E(Y_i)$. La distribution conjointe admet alors pour densité :

$$f(\mathbf{y}, \mu) = \prod_{i=1}^{I} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

La somme $N(N = y_+ = \sum_i y_i)$ des I variables aléatoires de Poisson indépendantes est également une variable de Poisson de paramètre $\mu_+ = \sum_i \mu_i$.

Multinomiale

En pratique, le nombre total n d'observations est souvent fixé a priori par l'expérimentateur et ceci induit une contrainte sur la somme des y_i. La distribution conjointe des variables Y_i est alors conditionnée par n et la densité devient :

$$f(\mathbf{y}, \mu) = \prod_{i=1}^{I} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \Big/ \frac{\mu_+^n e^{-\mu_+}}{n!}.$$

Comme $\mu_+^n = \sum_i \mu_+^{y_i}$ et $e^{-\mu_+} = \prod_i e^{-\mu_i}$, en posant $\pi_i = \frac{\mu_i}{\mu_+}$, on obtient :

$$f(\mathbf{y}, \mu) = n! \prod_{i=1}^{I} \frac{\pi_i^{y_i}}{y_i!} \quad \text{avec} \quad \sum_{i=1}^{I} \pi_i = 1 \text{ et } 0 \leq \pi_i \leq 1; i = 1, I.$$

On vérifie donc que $f(\mathbf{y}, \mu)$ est la fonction de densité d'une loi multinomiale dans laquelle les paramètres π_i modélisent les probabilités d'occurrence associées à chaque cellule. Dans ce cas, $E(Y_i) = n\pi_i$.

Produit de multinomiales

Dans d'autres circonstances, des effectifs marginaux lignes, colonnes ou sous-tables, peuvent être également fixés par l'expérimentateur comme dans le cas d'un sondage stratifié. Cela correspond au cas où une ou plusieurs variables sont contrôlées et ont donc un rôle explicatif ; leurs modalités sont connues *a priori*. Les lois de chacun des sous-éléments de la table, conditionnées par l'effectif marginal correspondant sont multinomiales. La loi conjointe de l'ensemble est alors un produit de multinomiales.

Conséquence

Trois modèles de distribution : Poisson, multinomial, produit de multinomiales, sont envisageables pour modéliser Y_i en fonction des conditions expérimentales. D'un point de vue théorique, on montre que ces modèles conduisent aux mêmes estimations des paramètres par maximum de vraisemblance. La différence introduite par le conditionnement intervient par une contrainte qui impose la présence de certains paramètres dans le modèle, ceux reconstruisant les marges fixées.

3.3 Modèles à 2 variables

Soit une table de contingence $(J \times K)$ issue du croisement de deux variables qualitatives X^1 à J modalités et X^2 à K modalités et dont l'effectif total n est fixé. La loi conjointe des effectifs Y_{jk} de chaque cellule est une loi multinomiale de paramètre π_{jk} et d'espérance :

$$E(Y_{jk}) = n\pi_{jk}.$$

Par définition, les variables X^1 et X^2 sont *indépendantes* si et seulement si :

$$\pi_{jk} = \pi_{+k}\pi_{j+}$$

où π_{j+} (resp. π_{+k}) désigne la loi marginale de X^1 (resp. X^2) :

$$\pi_{j+} = \sum_{k=1}^{K} \pi_{jk} \quad \text{et} \quad \pi_{+k} = \sum_{j=1}^{J} \pi_{jk}.$$

Si l'indépendance n'est pas vérifiée, on peut décomposer :

$$E(Y_{jk}) = n\pi_{jk} = n\pi_{j+}\pi_{+k}\frac{\pi_{jk}}{\pi_{j+}\pi_{+k}}.$$

Notons $\eta_{jk} = \ln(E(Y_{jk}))$. L'intervention de la fonction logarithme permet de linéariser la décomposition précédente autour du "modèle d'indépendance" :

$$\eta_{jk} = \ln n + \ln \pi_{j+} + \ln \pi_{+k} + \ln\left(\frac{\pi_{jk}}{\pi_{j+}\pi_{+k}}\right).$$

Ce modèle est dit *saturé* car, présentant autant de paramètres que de données, il explique exactement celles-ci. L'indépendance est vérifiée si le dernier terme de cette expression, exprimant une dépendance ou interaction comme dans le modèle d'analyse de variance, est nul pour tout couple (j, k).

Les logiciels mettent en place d'autres paramétrisations en faisant apparaître des effets différentiels, soit par rapport à une moyenne, soit par rapport à la dernière modalité.

Dans le premier cas, en posant :

$$\begin{aligned}
\beta_0 &= \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\eta_{jk} = \eta_{..},\\
\beta_j^1 &= \frac{1}{K}\sum_{k=1}^{K}\eta_{jk} - \eta_{..} = \eta_{j.} - \eta_{..},\\
\beta_k^2 &= \frac{1}{J}\sum_{j=1}^{J}\eta_{jk} - \eta_{..} = \eta_{.k} - \eta_{..},\\
\beta_{jk}^{12} &= \eta_{jk} - \eta_{j.} - \eta_{.k} + \eta_{..},
\end{aligned}$$

avec les relations :

$$\forall j, \forall k, \sum_{j=1}^{J}\beta_j^1 = \sum_{k=1}^{K}\beta_k^2 = \sum_{j=1}^{J}\beta_{jk}^{12} = \sum_{k=1}^{K}\beta_{jk}^{12} = 0,$$

le modèle saturé s'écrit :

$$\ln(E(Y_{jk})) = \eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_{jk}^{12}.$$

Il se met sous la forme matricielle

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

où $\mathbf{X}$ est la matrice expérimentale (design matrix) contenant les indicatrices. L'indépendance est obtenue lorsque tous les termes d'interaction β_{jk}^{12} sont nuls.

La deuxième paramétrisation considère la décomposition :

$$\pi_{jk} = \pi_{JK} \frac{\pi_{Jk}}{\pi_{JK}} \frac{\pi_{jK}}{\pi_{JK}} \frac{\pi_{jk}\pi_{JK}}{\pi_{Jk}\pi_{jK}}.$$

En posant :

$$\begin{aligned}
\beta_0 &= \ln n + \ln \pi_{JK}, \\
\beta_j^1 &= \ln \pi_{jK} - \ln \pi_{JK}, \\
\beta_k^2 &= \ln \pi_{Jk} - \ln \pi_{JK}, \\
\beta_{jk}^{12} &= \ln \pi_{jk} - \ln \pi_{jK} - \ln \pi_{Jk} + \ln \pi_{JK},
\end{aligned}$$

avec les mêmes relations entre les paramètres. Le modèle se met encore sous la forme :

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

et se ramène à l'indépendance si tous les paramètres β_{jk}^{12} sont nuls.

Si l'hypothèse d'indépendance est vérifiée, on peut encore analyser les effets principaux :

$$\text{si}, \forall j, \beta_j^1 = 0 \quad \text{alors,} \quad \pi_{jk} = \pi_{Jk} = \frac{1}{J}\pi_{+k}.$$

Il y a équiprobabilité des modalités de X^1. Même chose avec X^2 si les termes β_k^2 sont tous nuls.

Les paramètres du modèle log-linéaire sont estimés en maximisant la log-vraisemblance dont l'explicitation est reportée au chapitre suivant comme cas particulier de modèle linéaire généralisé. Pour les modèles simples, les estimations sont déduites des effectifs marginaux mais comme, dès que le modèle est plus compliqué, des méthodes itératives sont nécessaires, elles sont systématiquement mises en œuvre.

3.4 Modèle à trois variables

On considère une table de contingence $(J \times K \times L)$ obtenue par croisement de trois variables qualitatives X^1, X^2, X^3. La définition des paramètres est conduite de manière analogue au cas de deux variables en faisant apparaître des effets principaux et des interactions. Le modèle saturé se met sous la forme :

$$\ln(E(Y_{jkl})) = \eta_{jkl} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23} + \beta_{jkl}^{123}$$

et peut aussi est présenté sous forme matricielle.

Nous allons expliciter les sous-modèles obtenus par nullité de certains paramètres et qui correspondent à des structures particulières d'indépendance. Une façon classique de nommer les modèles consiste à ne citer que les interactions retenues les plus complexes. Les autres, ainsi que les effets principaux, sont contenues de par la structure hiérarchique du modèle. Ainsi, le modèle saturé est désigné par $(X^1X^2X^3)$ correspondant à la syntaxe `X1|X2|X3` de SAS.

Cas poissonnien ou multinomial

Seul le nombre total d'observations n est fixé dans le cas multinomial, ceci impose simplement la présence de β_0 dans le modèle.

i. Modèle partiel d'association ou de tout interaction d'ordre 2 : (X^1X^2, X^2X^3, X^1X^3)
Les termes β_{jkl}^{123} sont tous nuls, seules les interactions d'ordre 2 sont présentes. C'est le modèle implicitement considéré par l'analyse multiple des correspondances. Il s'écrit :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23}.$$

ii. Indépendance conditionnelle : (X^1X^2, X^1X^3)
Si, en plus, l'un des termes d'interaction est nul, par exemple $\beta_{kl} = 0$ pour tout couple (k, l), on dit que X^2 et X^3 sont indépendantes conditionnellement à X^1 et le modèle devient :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13}.$$

iii. Variable indépendante : (X^1, X^2X^3)
Si deux termes d'interaction sont nuls : $\beta_{jl}\beta_{jk} = 0$ pour tout triplet (j, k, l), alors X^1 est indépendante de X^2 et X^3.
$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{23}.$$

iv. Indépendance : (X^1, X^2, X^3)
Tous les termes d'interaction sont nuls :
$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3$$
et les variables sont mutuellement indépendantes.

Produit de multinomiales

- Si une variable est explicative, par exemple X^3, ses marges sont fixées, le modèle doit nécessairement conserver les paramètres
$$\eta_{jk} = \beta_0 + \beta_l^3 + \cdots$$
- Si deux variables sont explicatives, par exemple X^2 et X^3, le modèle doit conserver les termes :
$$\eta_{jk} = \beta_0 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{23} + \cdots$$

La généralisation à plus de trois variables ne pose pas de problème théorique. Les difficultés viennent de l'explosion combinatoire du nombre de termes d'interaction et de la complexité des structures d'indépendance. D'autre part, si le nombre de variables est grand, on est souvent confronté à des tables de contingence creuses (beaucoup de cellules vides) qui rendent défaillant le modèle log-linéaire. Une étude exploratoire (correspondances multiples par exemple) préalable est nécessaire afin de réduire le nombre des variables considérées et celui de leurs modalités.

4 Choix de modèle

4.1 Recherche pas à pas

Principalement deux critères (test du rapport de vraisemblance et test de Wald), décrits dans le chapitre suivant pour un cadre plus général, sont utilisés. Ces critères sont utilisés comme le test de Fisher du modèle linéaire gaussien. Ils permettent de comparer un modèle avec un sous-modèle et d'évaluer l'intérêt de la présence des termes complémentaires. On suit ainsi une stratégie descendante à partir du modèle complet ou saturé dans le cas du modèle log-linéaire. L'idée est de supprimer, un terme à la fois, la composante d'interaction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald. Les tests présentent une structure hiérarchisée. SAS facilite cette recherche en produisant une décomposition (Type III) de ces indices permettant de comparer chacun des sous-modèles excluant un des termes avec le modèle les incluant tous.

Attention, du fait de l'utilisation d'une transformation non linéaire (logit), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction. Enfin, selon les conditions expérimentales qui peuvent fixer les marges d'une table de contingence, la présence de certains paramètres est imposée dans un modèle log-linéaire.

4.2 Validation croisée

Pour le cas de la régression logistique, d'autres démarches plus calculatoires sont mises en œuvre pour comparer ou évaluer des modèles. Disposant de deux échantillons : un échantillon d'apprentissage et un de test, des modèles sont estimés avec l'échantillon d'apprentissage et comparés au regard de leur performance sur l'échantillon test.

Soit z_i la valeur de la variable binaire pour la $i$ème observation.

$$\begin{array}{lll} \text{si } \hat{\pi}_i > 0.5 & \text{alors} & \hat{z}_i = 1, \\ & \text{sinon} & \hat{z}_i = 0. \end{array}$$

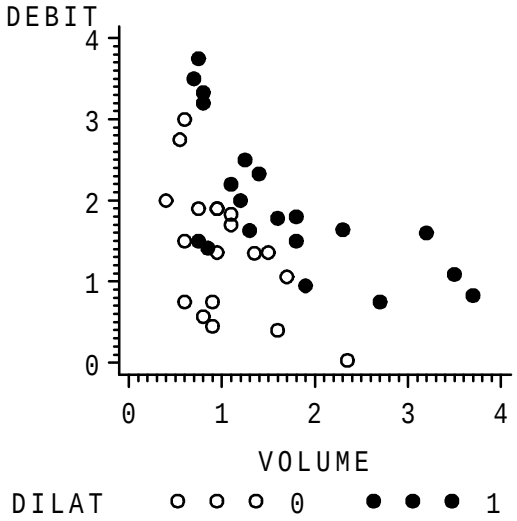


FIG. 4.1 – Nuage des modalités de Y dans les coordonnées des variables explicatives.

Un individu est dit *bien classé* si $\hat{z}_i = z_i$ et les modèles sont comparés à l'aide de leur pourcentage de bien classés sur l'échantillon test.

Si le nombre d'observations disponibles est trop faible, on peut systématiser cette démarche en éliminant une à une chaque observation pour construire n échantillons d'apprentissages tandis que les observations éliminées servent successivement de test.

5 Exemples

5.1 Modèle binomial

Il y a au moins 5 façons différentes d'estimer une régression logistique avec SAS :

`sas/logistic` cette procédure n'est utilisable que lorsque toutes les variables explicatives sont quantitatives, elle inclut en option un algorithme de recherche de modèles par sélection ascendante, descendante ou pas à pas.

`sas/catmod` est adaptée à toute modélisation impliquant des variables qualitatives (modèle log-linéaire, logit, probit...) mais la paramétrisation mise en œuvre rend les résultats difficiles à interpréter.

`sas/genmod` Cette procédure plus récente est arrivée avec la version 6.09. Elles est directement issue des travaux sur le modèle linéaire généralisé (Mc Cullagh et Nelder 1983) et s'inscrit donc dans la logique du logiciel GLIM pour la définition des modèles.

`sas/insight` est, pour le modèle linéaire généralisé, une version interactive et graphique de `sas/genmod`.

Enfin, pour mémoire, il existe de plus une option `logit` dans la procédure `sas/probit`. On se propose de comparer les résultats sur différents jeux de données.

Débits×Volumes

On étudie l'influence du débit et du volume d'air inspiré sur l'occurence (codée 1) de la dilatation des vaisseaux sanguins superficiels des membres inférieurs.

Référence : Pregibon, D. (1981) Logistic regression diagnostics, Annals of Stat., 9, 705-724.

Un graphique élémentaire représentant les modalités de Y dans les coordonnées de $X^1 \times X^2$ est toujours instructif. Il montre une séparation raisonnable et de bon augure des deux nuages de points. Dans le cas de nombreuses variables explicatives quantitatives, une analyse en composantes principales s'impose. Les formes des nuages représentés, ainsi que l'allure des distributions (étudiées préalablement), incitent dans ce cas à considérer par la suite les logarithmes des variables. Une variable (un) ne contenant que des "1"

dénombrant le nombre d'essais est nécessaire dans la syntaxe de genmod. Les données sont en effet non groupées.

Programmes et résultats :

```
proc logistic data=sasuser.debvol;
model dilat=l_debit l_volume;
run;
proc genmod data=sasuser.debvol;
model dilat/un=l_debit l_volume/d=bin;
run;
```

```
The LOGISTIC Procedure
                            Intercept
                 Intercept      and
 Criterion         Only      Covariates    Chi-Square for Covariates
 AIC                56.040       35.216          .
 SC                 57.703       40.206          .
 -2 LOG L           54.040       29.216(1)    24.824 with 2 DF (p=0.0001)
 Score                .            .          16.635 with 2 DF (p=0.0002)

            Parameter(2) Standard    Wald(3)    Pr >    Standardized  Odds
Variable DF  Estimate    Error   Chi-Square  Chi-Square  Estimate     Ratio
INTERCPT 1     2.8782    1.3214     4.7443     0.0294           .      .
L_DEBIT  1    -4.5649    1.8384     6.1653     0.0130    -2.085068    0.010
L_VOLUME 1    -5.1796    1.8653     7.7105     0.0055    -1.535372    0.006

       Association of Predicted Probabilities and Observed Responses(4)
                Concordant = 93.7          Somers' D = 0.874
                Discordant =  6.3          Gamma     = 0.874
                Tied       =  0.0          Tau-a     = 0.448
                (380 pairs)                c         = 0.937
```

Cette procédure fournit des critères de choix de modèle dont la déviance (1), le vecteur **b** des paramètres (2) et les statistiques des tests (3) comparant le modèle excluant un terme par rapport au modèle complet tel qu'il est décrit dans la commande.

```
The GENMOD Procedure
              Criteria For Assessing Goodness Of Fit
            Criterion             DF         Value      Value/DF
            Deviance              36       29.2156        0.8115 (1)
            Scaled Deviance       36       29.2156        0.8115 (2)
            Pearson Chi-Square    36       34.2516        0.9514 (3)
            Scaled Pearson X2     36       34.2516        0.9514
            Log Likelihood         .      -14.6078             .

                    Analysis Of Parameter Estimates
        Parameter    DF    Estimate (4)  Std Err   ChiSquare (5) Pr>Chi
        INTERCEPT     1     -2.8782      1.3214      4.7443      0.0294
        L_DEBIT       1      4.5649      1.8384      6.1653      0.0130
        L_VOLUME      1      5.1796      1.8653      7.7105      0.0055
        SCALE (6)     0      1.0000      0.0000           .       .
```

(1) Déviance du modèle par rapport au modèle saturé.
(2) Déviance pondérée si le paramètre d'échelle est différent de 1 en cas de sur-dispersion.
(3) Statistique de Pearson, voisine de la déviance, comparant le modèle au modèle saturé .
(4) Paramètres du modèle.
(5) Statistique des tests comparant le modèle excluant un terme par rapport au modèle complet.
(6) Estimation du paramètre d'échelle si la quasi-vraisemblance est utilisée.

Survie de poissons

On observe le nombre (Y parmi N) de décès de deux espèces de poissons en fonction de différentes valeurs de la température de l'eau. Les données sont restructurées afin de faire apparaître dans une autre

table la variable dichotomique (suc) codant le décès ou la survie ainsi que la variable (effet) exprimant le nombre d'occurence de chaque situations. Cette présentation est imposée par la procédure `catmod`.

```
proc catmod data=sasuser.poisson;
weight effect;
direct temp;
model suc=temp espece espece*temp;
run;
proc genmod data=sasuser.poissonr;
class espece;
model y/n=temp espece espece*temp/ dist=bin;
run;
```

```
CATMOD PROCEDURE
             MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE
            Source                    DF   Chi-Square      Prob
            ---------------------------------------------------
            INTERCEPT                  1        96.11    0.0000
            TEMP                       1        95.57    0.0000
            ESPECE                     1         0.02    0.8868
            TEMP*ESPECE                1         0.38    0.5403
            LIKELIHOOD RATIO          10         3.43    0.9694

                ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES
                                              Standard    Chi-
      Effect            Parameter  Estimate    Error    Square   Prob
      ------------------------------------------------------------------
      INTERCEPT                 1  -12.0110    1.2251    96.11  0.0000
      TEMP                      2    0.4931    0.0504    95.57  0.0000
      ESPECE                    3   -0.1745    1.2251     0.02  0.8868
      TEMP*ESPECE               4   -0.0309    0.0504     0.38  0.5403

GENMOD Procedure
                 Criteria For Assessing Goodness Of Fit
            Criterion             DF         Value      Value/DF
            Deviance              10        3.4297        0.3430
            Scaled Deviance       10        3.4297        0.3430
            Pearson Chi-Square    10        3.2340        0.3234
            Scaled Pearson X2     10        3.2340        0.3234
            Log Likelihood         .     -117.8030             .

                     Analysis Of Parameter Estimates
      Parameter          DF    Estimate     Std Err   ChiSquare  Pr>Chi
      INTERCEPT           1    -11.8366      1.8606     40.4704  0.0001
      TEMP                1      0.5240      0.0800     42.8793  0.0001
      ESPECE       1      1     -0.3489      2.4503      0.0203  0.8868
      ESPECE       2      0      0.0000      0.0000           .       .
      TEMP*ESPECE  1      1     -0.0618      0.1009      0.3750  0.5403
      TEMP*ESPECE  2      0      0.0000      0.0000           .       .
      SCALE                0      1.0000      0.0000           .       .
```

Les deux procédures produisent sur cet exemple les mêmes types de résultats : déviance par rapport au modèle saturé, décomposition de la vraisemblance et tests sur la présence des termes dans le modèle. La différence majeure apparaît dans la paramétrisation utilisée qui conditionne les valeurs des estimations. Néanmoins les modèles sont identiques. On peut s'en assurer en explicitant le modèle pour chaque espèce de poisson. Pour `catmod` espèce 1 et espèce 2 sont respectivement paramétrées $+1$ et -1 tandis que `genmod` utilise 1 et 0.

5.2 Modèle poissonien

On étudie les résultats d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province de l'Alberta à Edmonton au Canada (Jobson, 1991). Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table croisant :

i. Etat du conducteur : Normal ou Alcoolisé

ii. Port de la ceinture : Oui Non

iii. Gravité des blessures : 0 : rien à 3 : fatales

La procédure `genmod` est utilisée :

```
proc genmod data=sasuser.ceinture;
class co ce b ;
model effectif=co|ce|b @2  /type3 obstats dist=poisson;
run;
```

Une extraction des résultats donnent :

```
                Criteria For Assessing Goodness Of Fit
          Criterion              DF         Value      Value/DF
          Deviance                3        5.0136        1.6712

                  LR Statistics For Type 3 Analysis
                  Source       DF   ChiSquare  Pr>Chi

                  CO            1   3431.0877  0.0001
                  CE            1   3041.5499  0.0001
                  CO*CE         1    377.0042  0.0001
                  B             3  28282.8778  0.0001
                  CO*B          3    474.7162  0.0001
                  CE*B          3     42.3170  0.0001

                    Analysis Of Parameter Estimates
    Parameter           DF    Estimate     Std Err   ChiSquare  Pr>Chi
    INTERCEPT            1      3.6341      0.1550    550.0570  0.0001
    CO         A         1     -2.2152      0.1438    237.3628  0.0001
    CE         N         1      1.8345      0.1655    122.8289  0.0001
    CO*CE      A  N      1      0.9343      0.0545    293.9236  0.0001
    B          0         1      5.7991      0.1552   1396.7752  0.0001
    B          1         1      2.7848      0.1598    303.6298  0.0001
    B          2         1      2.1884      0.1637    178.7983  0.0001
    CO*B       A  0      1     -1.4622      0.1354    116.5900  0.0001
    CO*B       A  1      1     -0.6872      0.1423     23.3154  0.0001
    CO*B       A  2      1     -0.5535      0.1452     14.5293  0.0001
    CE*B       N  0      1     -0.2333      0.1658      1.9807  0.1593
    CE*B       N  1      1     -0.0902      0.1708      0.2786  0.5976
    CE*B       N  2      1      0.0741      0.1748      0.1799  0.6715

                         Observation Statistics
EFFECTIF       Pred      Xbeta        Std    HessWgt      Lower      Upper
   12500      12497     9.4332   0.008930      12497      12280      12718
     604   613.3370     6.4189     0.0395   613.3370   567.6707   662.6770
     344   337.8089     5.8225     0.0530   337.8089   304.5010   374.7601
      38    37.8677     3.6341     0.1550    37.8677    27.9495    51.3053
   61971      61974    11.0345   0.004016      61974      61488      62464
   ...
```

Les résultats montrent que le modèle de toute interaction d'ordre 2 est acceptable (déviance) et il semble que tous les termes soient nécessaires, toutes les interactions doivent être présentes au sens du test de Wald.

6 Exercices

Exo1

i. Expliciter la log-vraisemblance d'un échantillon de I observations de variables binomiales de paramètres n_i et π_i.

ii. Soit $\mathbf{X}$ une matrice de plan d'expérience regroupant l'observation de p variables explicatives. Écrire la log-vraisemblance du modèle de régression logistique associé et son expression (en fonction de $\hat{\pi}_i$) après maximisation.

iii. Exprimer la log-vraisemblance du modèle saturé en fonction de $\tilde{\pi}_i = y_i/n_i$ et la déviance du modèle.

iv. Exprimer la déviance en fonction du nombre de succès estimés ($\hat{y}_i = n_i\hat{\pi}_i$). Que devient la déviance si $n_i = 1$ (données non groupées) ?

v. Dériver la log-vraisemblance du modèle et en déduire les équations.

vi. On se place dans le cas particulier d'une seule variable explicative X binaire (0,1) et sans interaction. Montrer que l'estimateur du M.V. du coefficient associé à cette variable dans la régression logistique est le log de son odds ratio avec la variable Y.

Exo2

Les données (Jobson 1992) étudiées dans cet exercice sont issues d'une enquête réalisée auprès de 200 femmes mariées du Michigan. Les variables considérées sont les suivantes : THISYR, la variable à expliquer, (1) si la femme travaille l'année en cours, (0) sinon ; CHILD1 code la présence (1) ou l'absence (0) d'un enfant de moins de 2 ans ; CHILD2 présence ou absence d'un enfant entre 2 et 6 ans ; BLACK l'ascendance noire (1) ou blanche (0) ; les autres variables, âge (AGE), nombre d'années d'études (EDUC), revenu du mari (HUBINC) sont quantitatives.

i. On s'intéresse d'abord à expliquer la variable THISYR par la variable CHILD1. La table de contingence croisant ces deux variables sur les 100 premières observations est :

		CHILD1 = 0	CHILD1 = 1
THISYR	0	23	5
	1	71	1

Compléter la sortie SAS ci-dessous par les estimations des paramètres du modèle de régression logistique expliquant THISYR par CHILD1 sur ces mêmes 100 observations.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	??????	1.0954	2.1586	0.1418
CHILD1 0	1	??????	1.1214	5.9553	0.0147
CHILD1 1	0	??????	0.0000	.	.
SCALE	0	1.0000	0.0000	.	.

ii. On se propose de rechercher un "meilleur" modèle prédictif de la variable THISYR à partir des variables explicatives et de leurs éventuelles interactions par une méthode descendante, avec un seuil à 5%, basée soit sur le test de Wald soit sur celui du rapport de vraisemblance. Les tableaux ci-après, identifiées de A à E, constituent des étapes (dans le désordre) de ces recherches.

A — Type III (LR) Tests

Source	Chi-Sq	Pr > Chi-Sq
HUBINC	8.2651	0.0040
AGE	3.1370	0.0765
EDUC	5.9425	0.0148
BLACK	3.2647	0.0708
CHILD1	12.8633	0.0003
CHILD2	3.4808	0.0621

B — Type III (Wald) Tests

SourceDF	Chi-Sq	Pr > Chi-Sq
HUBINC	4.9989	0.0254
CHILD1	8.8465	0.0029
CHILD2	6.8243	0.0090

C — Type III (LR) Tests

Source	Chi-Sq	Pr > Chi-Sq
HUBINC	8.3902	0.0038
AGE	9.9545	0.0016
EDUC	6.0457	0.0139
BLACK	2.7937	0.0946
CHILD1	11.0479	0.0009
AGE*BLACK	3.9512	0.0468

D — Type III (Wald) Tests

Source	Chi-Sq	Pr > Chi-Sq
HUBINC	0.4228	0.5155
AGE	1.2265	0.2681
EDUC	3.0509	0.0807
BLACK	1.1083	0.2925
CHILD1	0.0050	0.9439
CHILD2	0.2103	0.6465
HUBINC*BLACK	0.7685	0.3807
HUBINC*CHILD1	0.2628	0.6082
HUBINC*CHILD2	1.9473	0.1629
AGE*BLACK	1.0663	0.3018
EDUC*CHILD2	0.3245	0.5689
BLACK*CHILD1	0.4325	0.5108

E — Type III (LR) Tests

Source	Chi-Sq	Pr > Chi-Sq
HUBINC	7.6979	0.0055
AGE	6.9126	0.0086
EDUC	6.5164	0.0107
BLACK	2.6241	0.1052
CHILD1	13.7913	0.0002
CHILD2	3.3785	0.0661
AGE*BLACK	3.8489	0.0498

Commenter chacun de ces tableaux en précisant s'il s'agit d'une étape intermédiaire (indiquer quelle est l'étape suivante), une étape finale, une erreur de sélection. Déduire de ces résultats les modèles finalement retenus pour chacune des stratégies de test (Wald, Vraisemblance).

iii. Pour le modèle issu du rapport de vraisemblance, interpréter le signe de chacun des paramètres associés aux effets principaux (résultats en annexe).

iv. Toujours pour ce modèle, expliciter les paramètres modèlisant le logit de la probabilité de travailler pour une femme blanche sans enfant et celui pour une femme noire également sans enfant. Comment interpréter l'interaction black*age ?

v. De son côté, Jobson (1992) retient le modèle explicatif considérant tous les effets principaux sans interaction (résultats en annexe). Sur le critère de la déviance des modèles retenus (Jobson, Wald, Vraisemblance), lequel vous semble meilleur ? Les capacités prédictives des modèles sont évaluées en comparant les prédictions et les observations de l'échantillon utilisé pour l'estimation (`APPRENTI`) puis celles de l'autre partie de l'échantillon (`TEST`) également de 100 personnes. Ces résultats sont présentés ci-dessous à l'aide de la procédure `freq` éditant la table de contingence croisant la variable observée et sa prédiction. Quel modèle retiendriez vous ?

```
        Modèle issu du              Modèle choisi          Modèle issu du test du
        test  de Wald                par Jobson(1992)      rapport de Vraisemblance

A       THISYR   PREDY              THISYR   PREDY             THISYR   PREDY
P      Frequency| 0| 1|Total      Frequency| 0| 1|Total      Frequency| 0| 1|Total
P      ---------+--+--+           ---------+--+--+           ---------+--+--+
R             0 |11|17| 28             0   |13|15| 28             0   |11|17| 28
E      ---------+--+--+           ---------+--+--+           ---------+--+--+
N             1 | 3|69| 72             1   | 5|67| 72             1   | 5|67| 72
T      ---------+--+--+           ---------+--+--+           ---------+--+--+
I      Total     14 86 100        Total     18 82 100        Total     16 84 100

        THISYR   PREDY              THISYR   PREDY             THISYR   PREDY
       Frequency| 0| 1|Total      Frequency| 0| 1|Total      Frequency| 0| 1|Total
T      ---------+--+--+           ---------+--+--+           ---------+--+--+
E             0 | 8|29| 37             0   |14|23| 37             0   |12|25| 37
S      ---------+--+--+           ---------+--+--+           ---------+--+--+
T             1 | 2|61| 63             1   |12|51| 63             1   | 6|57| 63
       ---------+--+--+           ---------+--+--+           ---------+--+--+
       Total     10 90 100        Total     26 74 100        Total     18 82 100
```

Annexe

```
*********
*** model thisyr/un= hubinc child1 child2 / d=bin type3 wald;
*********
          Criteria For Assessing Goodness Of Fit
    Criterion           Value      Value/DF
    Deviance              96       97.1842        1.0123
    Scaled Deviance       96       97.1842        1.0123
    Pearson Chi-Square    96      103.9713        1.0830
    Scaled Pearson X2     96      103.9713        1.0830
    Log Likelihood         .      -48.5921             .

  Analysis Of Parameter Estimates
Parameter       DF    Estimate     Std Err   ChiSquare  Pr>Chi
INTERCEPT        1     -1.0479      1.1296      0.8606  0.3536
HUBINC           1     -0.0822      0.0368      4.9989  0.0254
CHILD1     0     1      3.4648      1.1649      8.8465  0.0029
CHILD1     1     0      0.0000      0.0000           .       .
CHILD2           1     -1.3973      0.5349      6.8243  0.0090
SCALE            0      1.0000      0.0000           .       .

*********
*** model thisyr/un= age educ hubinc black child1 age*black / d=bin type3;
*********
   Criteria For Assessing Goodness Of Fit
     Criterion           Value      Value/DF
     Deviance              93       87.6180        0.9421
     Scaled Deviance       93       87.6180        0.9421
     Pearson Chi-Square    93       85.3283        0.9175
     Scaled Pearson X2     93       85.3283        0.9175
```

```
     Log Likelihood          .      -43.8090             .

      Analysis Of Parameter Estimates
Parameter       DF    Estimate     Std Err   ChiSquare  Pr>Chi
INTERCEPT        1    -21.6781      9.2962      5.4379  0.0197
AGE              1      0.4816      0.2850      2.8560  0.0910
EDUC             1      0.5237      0.2621      3.9932  0.0457
HUBINC           1     -0.1077      0.0443      5.9135  0.0150
BLACK      0     1     10.3427      8.0157      1.6649  0.1969
BLACK      1     0      0.0000      0.0000           .       .
CHILD1     0     1      4.2153      1.6416      6.5937  0.0102
CHILD1     1     0      0.0000      0.0000           .       .
AGE*BLACK  0     1     -0.4005      0.2873      1.9429  0.1634
AGE*BLACK  1     0      0.0000      0.0000           .       .
SCALE            0      1.0000      0.0000           .       .

*********
*** model thisyr/un= age educ hubinc black child1 child2 / d=bin type3;
*********
   Criteria For Assessing Goodness Of Fit
     Criterion           Value      Value/DF
     Deviance              93       88.0884        0.9472
     Scaled Deviance       93       88.0884        0.9472
     Pearson Chi-Square    93       96.6526        1.0393
     Scaled Pearson X2     93       96.6526        1.0393
     Log Likelihood         .      -44.0442             .

      Analysis Of Parameter Estimates
Parameter       DF    Estimate     Std Err   ChiSquare  Pr>Chi
INTERCEPT        1     -9.0352      3.7849      5.6986  0.0170
AGE              1      0.0773      0.0441      3.0766  0.0794
EDUC             1      0.4777      0.2292      4.3425  0.0372
HUBINC           1     -0.1079      0.0448      5.7929  0.0161
BLACK      0     1     -1.5451      0.9426      2.6867  0.1012
BLACK      1     0      0.0000      0.0000           .       .
CHILD1     0     1      4.5179      1.5736      8.2428  0.0041
CHILD1     1     0      0.0000      0.0000           .       .
CHILD2           1     -1.1238      0.6051      3.4491  0.0633
SCALE            0      1.0000      0.0000           .       .
```

Exo3

Soit trois variables qualitatives X^1, X^2 et Y à respectivement $J, K, 2$ modalités. Les observations sont rangées dans une table de contingence $I \times K \times 2$. On se propose de comparer le modèle de régression logistique expliquant Y par X^1 et X^2 (sans interaction) et le modèle log-linéaire (X^1X^2, X^1Y, X^2Y). Exprimer, dans la paramétrisation de SAS, les coefficients du modèle logistique en fonction de ceux du modèle log-linéaire. Vérifier avec les sorties SAS ci-dessous.

```
proc genmod data=sasuser.logitlog;
class X1 X2;
freq effectif;
model Y/un=X1 X2  / dist=bin; run;
```

```
       Parameter       DF    Estimate     Std Err   ChiSquare  Pr>Chi
       INTERCEPT        1      0.2718      0.1497      3.2955  0.0695
       X1         1     1     -0.6543      0.1742     14.1151  0.0002
       X1         2     1      0.9647      0.2139     20.3512  0.0001
       X1         3     0      0.0000      0.0000           .       .
       X2         1     1     -1.7596      0.2460     51.1764  0.0001
       X2         2     1      0.7703      0.2144     12.9122  0.0003
       X2         3     1     -0.2806      0.1915      2.1482  0.1427
       X2         4     0      0.0000      0.0000           .       .
       SCALE            0      1.0000      0.0000           .       .
```

```
proc genmod data=sasuser.logitlog;
class X1 X2 Y;
model effectif= X1|X2|Y @ 2 / dist=poi; run;
```

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	4.2698	0.1101	1503.2299	0.0001
X1	1		1	-0.3361	0.1557	4.6598	0.0309
X1	2		1	-0.4994	0.1741	8.2265	0.0041
X2	1		1	-2.5940	0.2814	84.9924	0.0001
X2	2		1	-0.3524	0.1662	4.4950	0.0340
X2	3		1	-0.9784	0.1870	27.3793	0.0001
X1*X2	1	1	1	0.5660	0.2647	4.5725	0.0325
X1*X2	1	2	1	0.2183	0.2146	1.0346	0.3091
X1*X2	1	3	1	0.2835	0.2208	1.6482	0.1992
X1*X2	2	1	1	2.0398	0.2956	47.6269	0.0001
X1*X2	2	2	1	-1.4646	0.4093	12.8021	0.0003
X1*X2	2	3	1	1.4366	0.2387	36.2149	0.0001
Y	0		1	-0.2718	0.1497	3.2955	0.0695
X1*Y	1	0	1	0.6543	0.1742	14.1151	0.0002
X1*Y	2	0	1	-0.9647	0.2139	20.3512	0.0001
X2*Y	1	0	1	1.7596	0.2460	51.1764	0.0001
X2*Y	2	0	1	-0.7703	0.2144	12.9122	0.0003
X2*Y	3	0	1	0.2806	0.1915	2.1482	0.1427

Chapitre 5

Introduction au modèle linéaire généralisé

L'objet de ce chapitre est d'introduire le cadre théorique global permettant de regrouper tous les modèles (linéaire gaussien, logit, log-linéaire) de ce cours et qui cherchent à exprimer l'espérance d'une variable réponse Y en fonction d'une combinaison linéaire des variables explicatives. Le *modèle linéaire généralisé* développé initialement en 1972 par Nelder et Wedderburn et dont on trouvera des exposés détaillés dans Nelder et Mc Cullagh (1983), Agresti (1990) ou Antoniadis et al. (1992), n'est ici qu'esquissé afin de définir les concepts communs à ces modèles : famille exponentielle, estimation par maximum de vraisemblance, tests, diagnostics, résidus. Il est mis en œuvre dans plusieurs logiciels dont GLIM, `glm` de Splus, `genmod` et `insight` de SAS.

1 Composantes des modèles

Les modèles catalogués dans la classe des modèles linéaires généralisés sont caractérisés par trois composantes.

1.1 Distribution

La *composante aléatoire* identifie la distribution de probabilités de la variable à expliquer. On suppose que l'échantillon statistique est constitué de n variables aléatoires $\{Y_i; i = 1, \ldots, n\}$ indépendantes admettant des distributions issues d'une *structure exponentielle*. Cela signifie que les lois de ces variables sont dominées par une même mesure dite de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme :

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi)\right\}. \tag{5.1}$$

Cette formulation inclut la plupart des lois usuelles comportant un ou deux paramètres : gaussienne, gaussienne inverse, gamma, Poisson, binomiale.... Le paramètre θ_i est appelé *paramètre naturel* de la famille exponentielle.

Attention, la mesure de référence change d'une structure exponentielle à l'autre, la mesure de Lebesgues pour une loi continue, une mesure discrète combinaison de masses de Dirac pour une loi discrète. Consulter Antoniadis et al. (1992) pour une présentation générale des structures exponentielles et des propriétés asymptotiques des estimateurs de leurs paramètres.

Pour certaines lois, la fonction u est de la forme :

$$u(\phi) = \frac{\phi}{\omega_i}$$

où les poids ω_i sont les poids connus des observations, fixés ici à 1 pour simplifier ; ϕ est appelé alors *paramètre de dispersion*, c'est un paramètre de nuisance intervenant, par exemple lorsque les variances des lois

gaussiennes sont inconnues, mais égal à 1 pour les lois à un paramètre (Poisson, binomiale). L'expression de la structure exponentielle (5.1) se met alors sous la *forme canonique* en posant :

$$\begin{aligned} Q(\theta) &= \frac{\theta}{\phi}, \\ a(\theta) &= \exp\left\{-\frac{v(\theta)}{\phi}\right\}, \\ b(y) &= \exp\{w(y,\phi)\}, \end{aligned}$$

on obtient

$$f(y_i,\theta_i) = a(\theta_i)b(y_i)\exp\left\{y_i Q(\theta_i)\right\}. \tag{5.2}$$

1.2 Prédicteur linéaire

Les observations planifiées des variables explicatives sont organisées dans la matrice $\mathbf{X}$ de planification d'expérience (design matrix). Soit $\boldsymbol{\beta}$ un vecteur de p paramètres, le prédicteur linéaire, *composante déterministe* du modèle, est le vecteur à n composantes :

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

1.3 Lien

La troisième composante exprime une *relation fonctionnelle* entre la composante aléatoire et le prédicteur linéaire. Soit $\{\mu_i = E(Y_i); i = 1,\ldots,n\}$, on pose

$$\eta_i = g(\mu_i) \quad i = 1,\ldots,n$$

où g, appelée *fonction lien*, est supposée monotone et différentiable. Ceci revient donc à écrire un modèle dans lequel une *fonction de la moyenne* appartient au sous-espace engendré par les variables explicatives :

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1,\ldots,n.$$

La fonction lien qui associe la moyenne μ_i au paramètre naturel est appelée *fonction lien canonique*. Dans ce cas,

$$g(\mu_i) = \theta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

1.4 Exemples

Loi gaussienne

Dans le cas d'un échantillon gaussien, les densités d'une famille de lois $\mathcal{N}(\mu_i,\sigma^2)$ s'écrit :

$$\begin{aligned} f(y_i,\mu_i) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\}\exp\left\{-\frac{1}{2}\frac{y_i^2}{\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)\right\}\exp\left\{y_i\frac{\mu_i}{\sigma^2}\right\} \end{aligned}$$

En posant

$$\begin{aligned} Q(\theta_i) &= \frac{\theta_i}{\phi} = \frac{\mu_i}{\sigma^2} \\ a(\theta_i) &= \exp\left\{-\frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\} \\ b(y_i) &= \exp\left\{-\frac{1}{2}\frac{y_i^2}{\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)\right\}. \end{aligned}$$

la famille gaussienne se met sous la forme canonique (5.2) qui en fait une famille exponentielle de paramètre de dispersion $\phi = \sigma^2$ et de paramètre naturel

$$\theta_i = E(Y_i) = \mu_i$$

et donc de fonction lien canonique, la fonction *identité*.

Loi de Bernouilli

Considérons n variables aléatoires binaires indépendantes Z_i de probabilité de succès π_i et donc d'espérance $E(Z_i) = \pi_i$. Les fonctions de densité de ces variables sont éléments de la famille :

$$f(z_i, \pi_i) = \pi_i^{z_i}(1-\pi_i)^{1-z_i} = (1-\pi_i)\exp\left\{z_i \ln \frac{\pi_i}{1-\pi_i}\right\},$$

qui est la forme canonique d'une structure exponentielle de paramètre naturel

$$\theta_i = \ln \frac{\pi_i}{1-\pi_i}.$$

Cette relation définit la fonction *logit* pour fonction lien canonique associée à ce modèle. La loi binomiale conduit à des résultats identiques en considérant les sommes de n_i (n_i connus) variables de Bernouilli.

Loi de Poisson

On considère n variables indépendantes Y_i de loi de Poisson de paramètre $\mu_i = E(Y_i)$. Les Y_i sont par exemple les effectifs d'une table de contingence. Ces variables admettent pour densités :

$$f(y_i, \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} = \exp\{-\mu_i\}\frac{1}{y_i!}\exp\{y_i \ln \mu_i\}$$

qui sont issues d'une structure exponentielle et, mises sous la forme canonique, de paramètre naturel

$$\theta_i = \ln \mu_i$$

définissant comme fonction lien canonique le *logarithme* pour ce modèle.

2 Estimation

L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé. Celle-ci s'exprime pour toute famille de distributions mise sous la forme (5.1) d'une structure exponentielle.

2.1 Expression des moments

Notons $\ell(\theta_i, \phi; y_i) = \ln f(y_i; \theta_i, \phi)$ la contribution de la $i$ème observation à la log-vraisemblance.

$$\ell(\theta_i, \phi; y_i) = [y_i\theta_i - v(\theta_i)]/u(\phi) + w(y_i, \phi).$$

L'étude du maximum de la log-vraisemblance nécessite la connaissance des dérivées :

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_i} &= [y_i - v'(\theta_i)]/u(\phi) \\ \frac{\partial^2 \ell}{\partial \theta_i^2} &= -v''(\theta_i)/u(\phi). \end{aligned}$$

Pour des lois issues de structures exponentielles, les conditions de régularité vérifiées permettent d'écrire :

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \quad \text{et} \quad -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2.$$

Alors,

$$E(Y_i) = \mu_i = v'(\theta_i)$$

et comme

$$E\{v''(\theta_i)/u(\phi)\} = E\{[Y_i - v'(\theta_i)]/u(\phi)\}^2 = \text{Var}(Y_i)/u^2(\phi)$$

il vient donc :

$$\text{Var}(Y_i) = v''(\theta_i)u(\phi)\ ;$$

justifiant ainsi l'appellation de *paramètre de dispersion* pour ϕ lorsque u est la fonction identité.

2.2 Équations de vraisemblance

Considérons p variables explicatives dont les observations sont rangées dans la matrice de plan d'expérience $\mathbf{X}$, $\boldsymbol{\beta}$ un vecteur de p paramètres et le prédicteur linéaire à n composantes

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

La fonction lien g est supposée monotone différentiable telle que : $\eta_i = g(\mu_i)$;
c'est la fonction lien canonique si : $g(\mu_i) = \theta_i$.

Pour n observations supposées indépendantes et en tenant compte que $\boldsymbol{\theta}$ dépend de $\boldsymbol{\beta}$, la log-vraisemblance s'écrit :

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \ell(\theta_i, \phi; y_i).$$

Calculons

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Comme

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= [y_i - v'(\theta_i)]/u(\phi) = (y_i - \mu_i)/u(\phi),\\
\frac{\partial \mu_i}{\partial \theta_i} &= v''(\theta_i) = \text{Var}(Y_i)/u(\phi),\\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \quad \text{car} \quad \eta_i = \mathbf{x}_i'\boldsymbol{\beta},\\
\frac{\partial \mu_i}{\partial \eta_i} & \quad \text{dépend de la fonction lien} \quad \eta_i = g(\mu_i),
\end{aligned}$$

Les équations de la vraisemblance sont :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0 \quad j = 1, \ldots, p.$$

Ce sont des équations non-linéaires en $\boldsymbol{\beta}$ dont la résolution requiert des méthodes itératives dans lesquelles interviennent le Hessien (pour Newton-Raphson) ou la *matrice d'information* (pour les Scores de Fisher). La matrice d'information est la matrice

$$\mathfrak{S} = \mathbf{X}'\mathbf{W}\mathbf{X}$$

de terme général

$$[\mathfrak{S}]_{jk} = E\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\text{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

et où $\mathbf{W}$ est la matrice diagonale de "pondération" :

$$[\mathbf{W}]_{ii} = \frac{1}{\text{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

2.3 Fonction lien canonique

Dans le cas particulier où la fonction lien du modèle linéaire généralisé utilisée est la fonction lien canonique associée à la structure exponentielle alors plusieurs simplifications interviennent :

$$\begin{aligned}
\eta_i &= \theta_i = \mathbf{x}_i'\boldsymbol{\beta},\\
\frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial v'(\theta_i)}{\partial \theta_i} = v''(\theta_i).
\end{aligned}$$

Ainsi,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} v''(\theta_i) x_{ij} = \frac{(y_i - \mu_i)}{u(\phi)} x_{ij}.$$

De plus, comme les termes $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}$ ne dépendent plus de y_i, on montre que le Hessien est égal à la matrice d'information et donc les méthodes de résolution du score de Fisher et de Newton-Raphson coïncident.

Si, de plus, $u(\phi)$ est constante pour les observations, les équations de vraisemblance deviennent :

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\boldsymbol{\mu}.$$

Ainsi, dans le cas gaussien, le modèle s'écrivant $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ avec la fonction de lien canonique identité, on retrouve la solution :

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

qui coïncide avec celle obtenue par minimisation des moindres carrés.

3 Qualité d'ajustement

Il s'agit d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations. Plusieurs critères sont proposés.

3.1 Déviance

Le modèle estimé est comparé avec le modèle dit *saturé*, c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Cette comparaison est basée sur l'expression de la *déviance* D des log-vraisemblances $\mathcal{L}$ et $\mathcal{L}_{\text{sat}}$:

$$D = -2(\mathcal{L} - \mathcal{L}_{\text{sat}})$$

qui est le logarithme du carré du rapport des vraisemblances. Ce rapport remplace ou "généralise" l'usage des sommes de carrés propres au cas gaussien et donc à l'estimation par moindres carrés.

On montre qu'asymptotiquement, D suit une loi du χ^2 à $n - p$ degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

Attention, l'approximation de la loi du χ^2 peut être douteuse. De plus, dans le cas de données non groupées (modèle binomial), le cadre asymptotique n'est plus adapté car le nombre de paramètres estimés tend également vers l'infini avec n et il ne faut plus se fier à ce test.

3.2 Test de Pearson

Un test du χ^2 est également utilisé pour comparer les valeurs observées y_i à leur prévision par le modèle. La statistique du test est définie par

$$X^2 = \sum_{i=1}^{I} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{\text{Var}}(\hat{\mu}_i)}$$

(μ_i est remplacé par $n_i \pi_i$ dans le cas binomial) et on montre qu'elle admet asymptotiquement la même loi que la déviance.

En pratique ces deux approches conduisent à des résultats peu différents et, dans le cas contraire, c'est une indication de mauvaise approximation de la loi asymptotique. Sachant que l'espérance d'une loi du χ^2 est son nombre de degrés de liberté et, connaissant les aspects approximatifs des tests construits, l'usage est souvent de comparer les statistiques avec le nombre de degrés de liberté. le modèle peut être jugé satisfaisant pour un rapport D/ddl plus petit que 1.

4 Tests

Deux critères sont habituellement proposés pour aider au choix de modèle.

4.1 Rapport de vraisemblance

Comme dans le cas de la régression multiple où un test permet de comparer un modèle avec un modèle réduit, le rapport de vraisemblance ou la différence de déviance est une évaluation de l'apport des variables explicatives supplémentaires dans l'ajustement du modèle. La différence des déviances entre deux modèles *emboîtés* respectivement à q_1 et q_2 $(q_2 > q_1)$ variables explicatives

$$\begin{aligned} D_2 - D_1 &= 2(\mathcal{L}_1 - \mathcal{L}_{\text{sat}}) - 2(\mathcal{L}_2 - \mathcal{L}_{\text{sat}}) \\ &= 2(\mathcal{L}_1 - \mathcal{L}_2) \end{aligned}$$

suit approximativement une loi du χ^2 à $(q_2 - q_1)$ degrés de liberté pour les lois à 1 paramètre (binomial, Poisson) et une loi de Fisher pour les lois à deux paramètres (gaussienne). Ceci permet donc de tester la significativité de la diminution de la déviance par l'ajout de variables explicatives ou la prise en compte d'interactions.

4.2 Test de Wald

Ce test est basé sur la forme quadratique faisant intervenir la matrice de covariance des paramètres, l'inverse de la matrice d'information observée $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$. Cette matrice est calculée à partir du Hessien approché par l'algorithme de maximisation. Elle généralise la matrice $(\mathbf{X}'\mathbf{X})^{-1}$ utilisée dans le cas du modèle linéaire gaussien en faisant intervenir une matrice $\mathbf{W}$ de pondération. Ainsi, test de Wald et test de Fisher sont équivalents dans le cas particulier du modèle gaussien.

Si la matrice $\mathbf{K}$, dite *contraste*, définit l'ensemble H_0 des hypothèses à tester sur les paramètres :

$$\mathbf{K}'\beta = 0,$$

on montre que la statistique

$$(\mathbf{K}'\mathbf{b})'(\mathbf{K}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}$$

suit asymptotiquement une loi du χ^2.

Attention, le test de Wald, approximatif, peut ne pas être précis si le nombre d'observations est faible.

5 Diagnostics

De nombreux indicateurs, comme dans le cas de la régression linéaire multiple, sont proposés afin d'évaluer la qualité ou la robustesse des modèles estimés. Ils concernent la détection des valeurs influentes et l'étude graphique des résidus. La définition de ces derniers pose quelques difficultés.

5.1 Effet levier

On construit la matrice de projection (hat matrix)

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}')\mathbf{W}^{1/2},$$

relative au produit scalaire de matrice $\mathbf{W}$, sur le sous-espace engendré par les variables explicatives. Les termes diagonaux de cette matrice supérieurs à $(3p/n)$ indiquent des valeurs potentiellement influentes. Le graphe représentant les points d'ordonnées h_{ii} et d'abscisses le numéro de l'observation les visualise.

5.2 Résidus

Avec des erreurs centrées, additives, c'est-à-dire dans le cas du modèle gaussien utilisant la fonction lien identité, il est naturel de définir des résidus par :

$$\varepsilon_i = y_i - E(y_i) = y_i - \mu_i.$$

comme dans le cas du modèle linéaire. Ce cadre est ici inadapté au cas général et différents substituts sont proposés. Chacun possède par ailleurs une version *standardisée* et une version *studentisée*.

Pearson

Les résidus obtenus en comparant valeurs observées y_i et valeurs prédites $\hat{y}_i$ sont pondérés par leur précision estimée par l'écart-type : s_i de $\hat{y}_i$. Ceci définit les résidus de Pearson :

$$r_{Pi} = \frac{y_i - \hat{y}_i}{s_i}$$

dont la somme des carrés conduit à la statistique du même nom. Ces résidus mesurent donc la contribution de chaque observation à la significativité du test découlant de cette statistique. Par analogie au modèle linéaire, on vérifie que ce sont également les résidus de la projection par la matrice $\mathbf{H}$.

Ces résidus ne sont pas de variance unité et sont donc difficiles à interpréter. Une estimation de leurs écarts-types conduit à la définition des résidus de Pearson standardisés :

$$r_{Psi} = \frac{y_i - \hat{y}_i}{s_i\sqrt{h_{ii}}}$$

faisant intervenir le terme diagonal de la matrice $\mathbf{H}$.

De plus, prenant en compte que les estimations des écarts-types s_i dépendent de la $i$ème observation et sont donc biaisés, des résidus studentisés sont obtenus en approchant au premier ordre le paramètre de dispersion $s_{(i)}$ calculé sans la $i$ème observation :

$$r_{Pti} = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{h_{ii}}}.$$

Déviance

Ces résidus mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Des versions standardisées et studentisées en sont définies comme pour ceux de Pearson.

Anscombe

Les lois des résidus précédents sont inconnues et même dissymétriques. Anscombe a donc proposé de faire opérer une transformation préalable afin de construire des résidus suivant une loi normale :

$$r_{Ai} = \frac{t(y_i) - t(\hat{y}_i)}{t'(y_i)s_i}.$$

L'explicitation de la fonction t dans le cadre du modèle linéaire généralisé est relativement complexe mais le calcul en est fourni par les logiciels. Comme précédemment, des versions standardisées et studentisées sont également calculées.

Un graphe utilisant ces résidus en ordonnées et les numéros d'observation en abscisses permet d'identifier les observations les moins bien ajustées par le modèle.

5.3 Mesure d'influence

De nombreux indicateurs sont proposés afin d'évaluer l'influence d'une observation sur l'estimation d'un paramètre, sur les prédictions ou encore sur la variance des estimateurs. Le plus utilisé, la distance de Cook, mesure globalement l'influence sur l'ensemble des paramètres. C'est la distance, au sens de la métrique définie par l'inverse de la covariance des paramètres, entre le vecteur des paramètres $\mathbf{b}$ estimé avec toutes les observations et celui estimé lorsque la $i$ème observation est supprimée.

$$D_i = \frac{1}{2}(\mathbf{b} - \mathbf{b}_{(i)})'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{b} - \mathbf{b}_{(i)}).$$

Cet indicateur prend simultanément en compte l'effet levier et l'importance du résidu de chaque observation. Le graphe de ces valeurs est donc plus synthétique et interprétable en tenant compte du graphe des résidus et de celui des termes diagonaux de $\mathbf{H}$.

6 Compléments

6.1 Sur-dispersion

Dans certaines situations, par exemple lors d'observations dépendantes, la variance de la variable Y_i supposée binomiale ou de Poisson, qui est théoriquement fixée par le modèle, est plus importante, multipliée par un facteur d'échelle (scale parameter) σ^2. Si ce paramètre est plus grand que 1, on dit qu'il y a sur-dispersion. Une méthode basée sur une maximisation de la formule de *quasi-vraisemblance* est alors utilisée pour estimer à la fois σ et β.

6.2 Variable "offset"

Lorsque la variable à expliquer dans le cas d'un modèle linéaire généralisé dépend également *linéairement* d'une autre variable, cette dernière est déclarée *offset* et sert ainsi à "tarer" le modèle. Exemple : pour modéliser le nombre de sinistres déclarés par catégorie de conducteurs, la variable *nombre de contrats* est déclarée "offset".

7 Exercices

Exo 1

Avec l'hypothèse valide dans le cas des familles exponentielles (densités deux fois différentiables), montrer les relations :

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \quad \text{et} \quad -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2.$$

Exo 2

i. On observe les réalisations y_i de n variables aléatoires indépendantes suivant des lois de Bernouilli de paramètre π_i. Soit $\mathbf{X}$ la matrice $(n \times (p+1))$ des observations issue de la planification expérimentale et contenant les variables explicatives. On s'intéresse au modèle de régression logistique :

$$\text{logit}(\pi_i) = x_i'\boldsymbol{\beta}$$

où $x_i \quad (1 \times (p+1))$ est le ième vecteur ligne de $\mathbf{X}$ mis en colonne et $\boldsymbol{\beta} \quad (1 \times (p+1))$ est le vecteur des paramètres. Écrire la log-vraisemblance $\mathcal{L}$ de l'échantillon $z_1, \ldots, z_n$ en fonction des paramètres $\boldsymbol{\beta}$ du modèle.

ii. On note $\mathbf{g} = \delta\mathcal{L}/\delta\boldsymbol{\beta}$ le gradient de $\mathcal{L}$; c'est le vecteur des dérivées $[\delta\mathcal{L}/\delta\beta_0, \ldots, \delta\mathcal{L}/\delta\beta_p]'$. Soit $\mathbf{z}$ le vecteur des observations et $\boldsymbol{\pi}$ celui des π_i. Montrer que le gradient se met sous la forme $\mathbf{g} = \mathbf{X}'\mathbf{z} - \mathbf{X}'\boldsymbol{\pi}$ (expliciter un terme général de ce vecteur).

iii. On note $\mathcal{H}$ le hessien de $\mathcal{L}$; c'est la matrice ayant pour terme général la dérivée seconde $\delta^2\mathcal{L}/\delta\beta_j\beta_k$. Montrer que $\mathcal{H}$ se met sous la forme $\mathcal{H} = -\mathbf{X}'\mathbf{W}\mathbf{X}$ et expliciter la matrice diagonale $\mathbf{W}$ en fonction des π_i.

iv. La procédure de Newton-Raphson est utilisée pour approcher le maximum de la fonction $\mathcal{L}(\mathbf{z}, \mathbf{X}, \boldsymbol{\beta})$ relativement à $\boldsymbol{\beta}$. Elle consiste à calculer par récurrence une séquence $\mathbf{b_n}$ d'estimateurs de $\boldsymbol{\beta}$ telle que $\mathcal{L}(\mathbf{z}, \mathbf{X}, \mathbf{b_{n+1}}) > \mathcal{L}(\mathbf{z}, \mathbf{X}, \mathbf{b_n})$. En notant respectivement $\mathbf{H}_n^{-1} = \mathcal{H}_{|\mathbf{b_n}}$, $\mathbf{d}_n$ et $\boldsymbol{\pi}_n$ les évaluations du hessien, du gradient et de $\boldsymbol{\pi}$ pour $\boldsymbol{\beta} = \mathbf{b_n}$, la récurrence s'écrit : $\mathbf{b_{n+1}} = \mathbf{b_n} - \mathbf{H_n g_n}$. Montrer que, dans le cas de la régression logistique sur variables de Bernouilli, cette récurrence s'écrit :

$$\mathbf{b_{n+1}} = \mathbf{b_n} + (\mathbf{X}'\mathbf{W_n}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{z} - \mathbf{X}'\boldsymbol{\pi_n})$$

et expliciter $\mathbf{W}_n$.

v. Suivre la même démarche en remplaçant le modèle précédent par un modèle poissonien avec la fonction lien canonique $\ln(\pi_i) = x_i'\boldsymbol{\beta}$. Que deviennent la log-vraisemblance,

vi. son gradient,

vii. son Hessien,

viii. la relation de récurrence ?

Bibliography

Agresti, A. (1990). *Categorical data analysis*. Wiley.

Antoniadis, A., Berruyer, J., and Carmona, R. (1992). *Régression non linéaire et applications*. Economica.

Collett, D. (1991). *Modelling binary data*. Chapman & Hall.

Dobson, A. (1990). *An introduction to generalized linear models*. Chapman and Hall.

Everitt, B. and Dunn, G. (1991). *Applied Multivariate Data Analysis*. Edward Arnold.

Jobson, J. (1991). *Applied Multivariate Data Analysis*, volume I : Regression and experimental design. Springer-Verlag.

Jobson, J. (1992). *Applied Multivariate Data Analysis*, volume II : Categorical and multivariate methods. Springer-Verlag.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*. Chapman & Hall.

Monfort, A. (1982). *Cours de Statistique Mathématique*. Economica.

SAS (1989). *SAS/STAT User's Guide*, volume 2. Sas Institute Inc., fourth edition. version 6.

SAS (1995). *SAS/INSIGHT User's Guide*. Sas Institute Inc., third edition. version 6.

Tomassonne, R., Audrain, S., Lesquoy-de Turckheim, E., and Millier, C. (1992). *La régression, nouveaux regards sur une ancienne méthode statistique*. Masson.