Université de Liège

Faculté de Médecine

Analyse statistique multivariée

Adelin Albert
Professeur ordinaire

Edition 2011

Préface

La statistique multivariée recouvre l'ensemble des méthodes d'analyse des données résultant de l'observation simultanée de plusieurs variables. Elle fait dès lors appel au calcul vectoriel et plus généralement au calcul matriciel. Il faut bien reconnaître que la plupart des problèmes rencontrés en pratique impliquent au moins deux variables et relèvent de ce fait du domaine multidimensionnel. L'exemple le plus célèbre est celui des iris de Fisher qui reprend les mesures de quatre variables, à savoir longueur et largeur des sépales, longueur et largeur des pétales, chez 50 iris setosa, 50 iris versicolor et 50 iris virginica. Ces données sont reprises en annexe. Elles permettent d'illustrer la plupart des méthodes statistiques multivariées décrites dans cet ouvrage.

Il n'est guère possible d'aborder la statistique multivariée sans avoir une connaissance approfondie des méthodes de l'analyse statistique univariée. Celles-ci sont décrites dans mon livre "Biostatistique". Les méthodes de l'analyse statistique multivariée ont été développées dans le courant du 20^{e} siècle mais n'ont pu être réellement exploitées qu'avec l'avènement de l'informatique. En fait, plus le nombre de variables étudiées simultanément est élevé, plus les calculs sont longs voire impossibles manuellement. Les ordinateurs actuels permettent de résoudre des problèmes multivariés complexes en quelques secondes. Par ailleurs, de nombreux logiciels ont été développés, rendant ainsi accessibles à une large communauté d'utilisateurs les techniques de l'analyse statistique multivariée. S'il s'agit là d'un progrès considérable, il est néanmoins impératif que l'utilisateur possède un minimum de connaissances en statistique multivariée. C'est l'objectif principal de cet ouvrage destiné aux étudiants et aux chercheurs.

Cet ouvrage est structuré en 8 chapitres. Le Chapitre 1 reprend les notions de base du calcul matriciel, outil indispensable à la présentation, à la caractérisation et à la résolution des problèmes statistiques multivariés. Au Chapitre 2, on rappelle brièvement les notions de population, d'échantillon ainsi que les différents types de variables (quantitatives, qualitatives et binaires). On montre ensuite comment se construit une matrice d'observations $n \times p$, tableau résultant de l'observation chez n sujets ou objets (lignes)

de p variables (colonnes). La représentation d'observations multivariées par différentes techniques est aussi abordée. Le Chapitre 3 étend au niveau multivarié les concepts classiques de moyenne et de variance. En particulier, on y introduit les notions de matrice de variances-covariances et de matrice de corrélations. La distance de Mahalanobis entre deux points de l'espace à p dimensions y est définie. Elle va au-delà de la distance euclidienne classique, de manière à tenir compte des associations entre les variables.

La première grande méthode de statistique multivariée est développée au Chapitre 4. Il s'agit de l'analyse en composantes principales qui permet de représenter dans un plan (espace à deux dimensions) la distribution d'un ensemble de points de l'espace multidimensionnel. On obtient ainsi une photographie de la matrice d'observations. On présente aussi brièvement dans ce chapitre la méthode plus récente dite du biplot.

Le Chapitre 5 s'intéresse à la relation et à la corrélation entre une variable dite "dépendante" et plusieurs autres variables dites "indépendantes". Il s'agit de la méthode de régression et de corrélation multiple. Dans la plupart des livres de statistique, cette méthode relève de la statistique univariée car les variables indépendantes sont considérées comme des facteurs fixés par l'utilisateur dans un plan d'expérience (plans factoriels) et seule la variable dépendante est observée. Nous l'avons reprise comme méthode statistique multivariée parce qu'elle fait appel au calcul matriciel mais aussi en raison du fait que, si les variables sont observées simultanément, il s'agit bien d'un problème multivarié.

Le Chapitre 6 fait référence à l'une des méthodes les plus utilisées actuellement en statistique multivariée. Il s'agit de la régression logistique qui permet d'étudier l'association entre une variable dépendante binaire et un vecteur de variables. On aborde également la méthode de régression logistique ordinale où, en lieu et place d'une variable binaire, on étudie une variable ordinale dont les catégories sont ordonnées.

Le Chapitre 7 est consacré aux durées de vie, sujet déjà abordé dans le Chapitre 5 du livre de Biostatistique. Plus spécifiquement, on s'intéresse ici à la relation entre une durée de vie et un ensemble de covariables par le biais de la méthode de régression de Cox, appelée aussi modèle des "risques proportionnels" de Cox. Il s'agit d'une méthode complexe mais de la plus haute actualité. Son utilisation dans la littérature internationale est abondante.

Enfin, le Chapitre 8 reprend un vieux problème de la statistique multivariée, celui de l'analyse discriminante. On se propose de séparer deux ou plusieurs populations sur base d'un vecteur de variables, grâce à la fonction linéaire discriminante de Fisher ou à son extension multiple. L'analyse discriminante canonique permet de représenter les populations sur un plan, à la manière de l'analyse en composantes principales. L'analyse discriminante

peut aussi être vue comme le problème de classement d'un sujet ou d'un objet dans deux ou plusieurs populations avec un risque minimum de se tromper. Cette approche est également abordée.

Les annexes reprennent trois fichiers de données servant à illustrer les méthodes décrites dans le livre : (1) les iris de Fisher, (2) les données des traumatisés crâniens, et (3) les données de patients atteints d'un cancer rectal. Les annexes contiennent aussi 4 des 7 tables figurant dans le livre de Biostatistique, à savoir les lois Normale, Chi-carré, t de Student et F de Snedecor.

Il existe aujourd'hui de nombreux logiciels permettant d'utiliser les méthodes de l'analyse statistique multivariée. Nous avons pris l'habitude d'utiliser SAS (SAS Institute, Cary, NC) et S-PLUS (MathSoft, Seattle, WA). Pour les étudiants, nous avons adopté le logiciel STATISTICA (Statsoft, Tulsa, OK), pour lequel un manuel d'utilisation a également été rédigé.

Au terme de ce travail, je tiens à remercier mon assistante Anne-Françoise Donneau qui m'a permis de finaliser ce livre débuté voici trop longtemps. Il sert de support didactique aux nombreux étudiants qui suivent le cours de même nom. J'exprime aussi ma reconnaissance à mes autres assistantes, Laurence Seidel et Sophie Vanbelle, pour leur contribution, ainsi qu'à Mme Bartholoméus et Mme Marchetta pour le travail de dactylographie.

Liège, mars 2006
Revu en juillet 2008
Revu en juillet 2011
Adelin Albert

Chapitre 1

Notions de calcul matriciel

1.1 Matrices

1.1.1 Définition

Une *matrice* $\underset{\sim}{A}$ de dimension $r \times c$ est un tableau rectangulaire à r lignes et c colonnes composé de nombres. Si a_{ij} désigne l'élément de la matrice $\underset{\sim}{A}$ à l'intersection de la i$^{\text{ème}}$ ligne et de la j$^{\text{ème}}$ colonne, la matrice complète peut s'écrire sous la forme

$$\underset{\sim}{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2c} \\ \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rc} \end{pmatrix}.$$

Lorsque $r = c$, on dit que la matrice est *carrée*. De plus, si $a_{ij} = a_{ji}$ quel que soit $i \neq j$, la matrice est dite *symétrique*. La trace d'une matrice carrée est la somme des éléments diagonaux de la matrice, $tr\underset{\sim}{A} = a_{11} + \ldots + a_{rr}$.

1.1.2 Vecteurs et scalaire

– Lorsque $c = 1$ (une seule colonne), la matrice se réduit à un *vecteur-colonne* et le second indice n'est plus nécessaire. Le vecteur s'écrit alors plus simplement

$$\underset{\sim}{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix}$$

et on dit qu'il est de dimension r.

– Lorsque $r = 1$ (une seule ligne), la matrice se réduit à un *vecteur-ligne* de dimension c que l'on note

$$\underset{\sim}{a}^T = (a_1, a_2, \ldots, a_c).$$

Comme, par définition, un vecteur est toujours un vecteur-colonne, on dit que le vecteur ci-dessus est *transposé*, d'où la notation particulière.

On peut donc dire qu'une matrice $\underset{\sim}{A}$ de dimension $r \times c$ est composée de c vecteurs-colonnes de dimension r ou de r vecteurs-lignes de dimension c

$$\underset{\sim}{A}_{r\times c} = (\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_c) = \begin{pmatrix} \underset{\sim}{a}_1^T \\ \vdots \\ \underset{\sim}{a}_r^T \end{pmatrix}.$$

– Lorsque $r = c = 1$, la matrice se réduit à un *scalaire* a_{11}. Donc, un scalaire est une matrice de dimension 1×1.

1.1.3 Exemples

La matrice 8×3 ci-dessous représente les observations chez 8 sujets de 3 variables, la taille (cm), le poids (kg) et la circonférence du bras (cm)

$$\begin{pmatrix} 167.9 & 71.8 & 30.0 \\ 183.8 & 75.1 & 29.4 \\ 172.9 & 58.0 & 26.0 \\ 175.5 & 58.4 & 25.7 \\ 176.4 & 67.7 & 27.9 \\ 168.5 & 75.2 & 31.7 \\ 178.0 & 67.3 & 27.4 \\ 178.0 & 71.3 & 29.0 \end{pmatrix}.$$

La matrice carrée symétrique 3×3 représente le tableau des corrélations entre les 3 variables

$$\begin{pmatrix} 1.0 & 0.049 & -0.30 \\ 0.049 & 1.0 & 0.93 \\ -0.30 & 0.93 & 1.0 \end{pmatrix}.$$

Le vecteur (167.9, 71.8, 30.0) représente les observations des trois variables chez le sujet N°1. C'est une matrice 1×3.

Enfin, la matrice 8×1 suivante

$$\begin{pmatrix} 71.8 \\ 75.1 \\ 58.0 \\ 58.4 \\ 67.7 \\ 75.2 \\ 67.3 \\ 71.3 \end{pmatrix}$$

représente le vecteur des observations de la variable N°2 (poids) chez les 8 sujets.

1.1.4 Matrices particulières

– La *transposée* de la matrice $\underset{\sim}{A}$ de dimension $r \times c$ est obtenue en transposant les lignes et les colonnes de $\underset{\sim}{A}$. Il s'agit d'une matrice de dimension $c \times r$, notée $\underset{\sim}{A}^T$. Ainsi, la transposée de la matrice 2×3

$$\underset{\sim}{A} = \begin{pmatrix} 167.9 & 71.8 & 30.0 \\ 183.8 & 75.1 & 29.4 \end{pmatrix}$$

est la matrice 3×2

$$\underset{\sim}{A}^T = \begin{pmatrix} 167.9 & 183.8 \\ 71.8 & 75.1 \\ 30.0 & 29.4 \end{pmatrix}.$$

On constate aisément qu'en transposant une matrice symétrique, on retrouve la matrice de départ. Donc, pour une matrice symétrique,

$$\underset{\sim}{A}^T = \underset{\sim}{A}$$

– Une *matrice diagonale* est une matrice symétrique dont tous les éléments sont nuls, à l'exception des éléments diagonaux, soit

$$\underset{\sim}{D} = \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ 0 & 0 & \dots & d_{rr} \end{pmatrix}$$

ou plus simplement

$$\text{diag}(d_1, d_2, \dots, d_r).$$

- La matrice *identité* de dimension $r \times r$ est la matrice diagonale dont les éléments diagonaux sont tous égaux à 1. On la note $\underset{\sim}{I}_r$ ou $\underset{\sim}{I}$. C'est la généralisation de l'unité "1".
- La matrice *nulle* est la matrice dont tous les éléments sont nuls, qu'elle soit rectangulaire ou carrée. On la note $\underset{\sim}{0}$.

1.2 Opérations sur les matrices

1.2.1 Addition

La somme de deux matrices $\underset{\sim}{A}$ et $\underset{\sim}{B}$ de même dimension $r \times c$ est une matrice $\underset{\sim}{C}$ de dimension $r \times c$

$$\underset{\sim}{A} + \underset{\sim}{B} = \underset{\sim}{C}$$

où $c_{ij} = a_{ij} + b_{ij}$.

1.2.2 Soustraction

La soustraction de deux matrices $\underset{\sim}{A}$ et $\underset{\sim}{B}$ de même dimension $r \times c$ est une matrice $\underset{\sim}{C}$ de dimension $r \times c$

$$\underset{\sim}{A} - \underset{\sim}{B} = \underset{\sim}{C}$$

où $c_{ij} = a_{ij} - b_{ij}$.

1.2.3 Multiplication par un scalaire

Si on multiplie une matrice $\underset{\sim}{A}$ de dimension $r \times c$ par un scalaire k, on obtient une matrice

$$\underset{\sim}{C} = k\underset{\sim}{A}$$

où $c_{ij} = ka_{ij}$. Tous les éléments de la matrice $\underset{\sim}{A}$ sont donc multipliés par le scalaire.

1.2.4 Produit de deux matrices

La produit de la matrice $\underset{\sim}{A}$ de dimension $r \times c$ par la matrice $\underset{\sim}{B}$ de dimension $r' \times c'$ requiert que le nombre de colonnes de la première soit égal au nombre de lignes de la seconde ($c = r'$). On obtient alors une matrice $\underset{\sim}{C}$

dont le nombre de lignes est celui de $\underset{\sim}{A}$ et le nombre de colonnes celui de $\underset{\sim}{B}$. En clair,

$$\underset{\sim}{A}_{r\times c} \cdot \underset{\sim}{B}_{c\times c'} = \underset{\sim}{C}_{r\times c'}$$

où $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{ic}b_{cj} = \sum_{k=1}^{c} a_{ik} \cdot b_{kj}$.

Le produit des deux matrices $\underset{\sim}{A}$ et $\underset{\sim}{B}$, que l'on note indistinctement $A.B, A \times B$ ou AB, s'effectue "lignes de $\underset{\sim}{A}$ par colonnes de $\underset{\sim}{B}$"; ainsi $c_{ij} = \underset{\sim}{a}_i^T \cdot \underset{\sim}{b}_j$.

A titre d'exemple, soient les deux matrices

$$\underset{\sim}{A}_{2\times 3} = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 5 & 1 \end{pmatrix} \quad \text{et} \quad \underset{\sim}{B}_{3\times 2} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 2 & 1 \end{pmatrix}.$$

Le produit $\underset{\sim}{A} \times \underset{\sim}{B}$ est possible et conduit à une matrice $\underset{\sim}{C}$ de dimension 2×2. Ainsi, l'élément c_{11} s'obtient en multipliant la première ligne de $\underset{\sim}{A}$ par la première colonne de $\underset{\sim}{B}$ et on a $c_{11} = 1 \times 1 + 0 \times 0 + 2 \times 2 = 5$. De même, $c_{12} = 1 \times 1 + 0 \times 0 + 2 \times 1 = 3$, $c_{21} = 3 \times 1 + 5 \times 0 + 1 \times 2 = 5$ et $c_{22} = 3 \times 1 + 5 \times 0 + 1 \times 1 = 4$.

Dès lors,

$$\underset{\sim}{C} = \underset{\sim}{A} \times \underset{\sim}{B} = \begin{pmatrix} 5 & 3 \\ 5 & 4 \end{pmatrix}.$$

Remarques

– Le produit de deux matrices carrées $\underset{\sim}{A}$ et $\underset{\sim}{B}$ de même dimension est toujours possible. Toutefois, en général, le produit matriciel n'est pas commutatif : $\underset{\sim}{A} \cdot \underset{\sim}{B} \neq \underset{\sim}{B} \cdot \underset{\sim}{A}$.

– Soit $\underset{\sim}{a}$ un vecteur de dimension r. Dans ce cas, le produit

$$\underset{\sim}{a}^T \cdot \underset{\sim}{a} = a_1^2 + a_2^2 + \ldots + a_r^2$$

est appelé *produit scalaire* car il donne un scalaire (appelé aussi *norme* de $\underset{\sim}{a}$ ou $\|\underset{\sim}{a}\|$), alors que le produit *matriciel*

$$\underset{\sim}{a} \cdot \underset{\sim}{a}^T = \begin{pmatrix} a_1^2 & a_1a_2 & \ldots & a_1a_r \\ a_2a_1 & a_2^2 & \ldots & a_2a_r \\ \ldots & \ldots & \ldots & \ldots \\ a_ra_1 & a_ra_2 & \ldots & a_r^2 \end{pmatrix}$$

conduit à une matrice symétrique de dimension $r \times r$.

– Lorsque le produit scalaire de deux vecteurs $\underset{\sim}{a}$ et $\underset{\sim}{b}$ de même longueur est nul, $\underset{\sim}{a}^T \underset{\sim}{b} = 0$, on dit que les vecteurs sont *orthogonaux*.

– Dans le cas d'une matrice carrée $\underset{\sim}{A}$, on vérifie aisément que

$$\underset{\sim}{A} \cdot \underset{\sim}{I} = \underset{\sim}{I} \cdot \underset{\sim}{A} = \underset{\sim}{A}$$

– Notons enfin que si $\underset{\sim}{A}$ est une matrice $r \times c$, le produit $\underset{\sim}{A} \cdot \underset{\sim}{A}^T$ est une matrice symétrique de dimension $r \times r$ et le produit $\underset{\sim}{A}^T \cdot \underset{\sim}{A}$ une matrice symétrique de dimension $c \times c$.

1.2.5 Généralisation

On peut à volonté additionner et soustraire des matrices de même dimension $r \times c$ ou en calculer des combinaisons linéaires telle que

$$k_1 \underset{\sim}{A}_1 + k_2 \underset{\sim}{A}_2 + \ldots + k_n \underset{\sim}{A}_n$$

où les k_i sont des scalaires.

Le produit de plusieurs matrices ne pose pas de problème pour autant que les matrices adjacentes répondent à la condition requise que le nombre de colonnes de chacune soit égal au nombre de lignes de celle qui la suit. Ainsi, le triple produit

$$\underset{\sim}{A}_{5\times3} \cdot \underset{\sim}{B}_{3\times8} \cdot \underset{\sim}{C}_{8\times7}$$

est possible et fournit comme résultat une matrice de dimension 5×7.

1.3 Inversion d'une matrice

1.3.1 Définition

L'inversion d'une matrice autorise en quelque sorte "la division" de matrices. Pour rappel, l'inverse d'un nombre $k \neq 0$ s'écrit $\frac{1}{k}$ ou k^{-1} et on a $k \cdot k^{-1} = 1$. En conséquence, si on divise deux nombres k_1 et k_2, cela revient à multiplier le premier par l'inverse du second, soit $k_1/k_2 = k_1 \times k_2^{-1}$. La division est donc un cas particulier de la multiplication.

L'inversion s'applique typiquement à une matrice carrée. Ainsi, l'inverse de la matrice carrée $\underset{\sim}{A}$ de dimension $r \times r$ est la matrice de même dimension, notée $\underset{\sim}{A}^{-1}$, telle que

$$\underset{\sim}{A} \cdot \underset{\sim}{A}^{-1} = \underset{\sim}{A}^{-1} \cdot \underset{\sim}{A} = \underset{\sim}{I}$$

Inverser une matrice n'est pas chose aisée lorsque la dimension de la matrice est élevée. L'inversion d'une matrice nécessite le calcul de son déterminant.

1.3.2 Déterminant

Le *déterminant* d'une matrice carrée $\underset{\sim}{A}$ de dimension $r \times r$ est le scalaire noté $dtm\underset{\sim}{A}$ ou $|\underset{\sim}{A}|$ et obtenu, quel que soit $1 \leq i \leq r$, par l'expression

$$|\underset{\sim}{A}| = a_{i1}A_{i1} + a_{i2}A_{i2} + \ldots + a_{ir}A_{ir}$$

où A_{ij} est le *cofacteur* associé à l'élément a_{ij}, c'est-à-dire la quantité

$$A_{ij} = (-1)^{i+j}\mathcal{M}_{ij}$$

où $\mathcal{M}_{ij}$ est le *mineur* associé à a_{ij}, c'est-à-dire le déterminant de la matrice obtenue en supprimant dans la matrice $\underset{\sim}{A}$ la i$^{\text{ème}}$ ligne et la j$^{\text{ème}}$ colonne.

Par convention, le déterminant d'un scalaire a est le scalaire lui-même : $|a| = a$.

A titre d'exemple, le déterminant de la matrice $\underset{\sim}{A}$ de dimension 2×2 vaut (en prenant $i = 1$)

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

De même, on vérifiera aisément que (en prenant $i = 1$)

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{33}a_{21}a_{12} - a_{32}a_{23}a_{11}.$$

1.3.3 Calcul de l'inverse

L'*inverse* de la matrice $\underset{\sim}{A}$ de dimension $r \times r$ est obtenue à partir de l'expression

$$\underset{\sim}{A}^{-1} = \frac{1}{|\underset{\sim}{A}|}\begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1r} \\ A_{21} & A_{22} & \ldots & A_{2r} \\ \ldots & \ldots & \ldots & \ldots \\ A_{r1} & A_{r2} & \ldots & A_{rr} \end{pmatrix}^T$$

où la matrice à droite est la matrice des cofacteurs mais transposée. On note parfois $a^{ij} = A_{ji}/|\underset{\sim}{A}|$ l'élément (i,j) de la matrice $\underset{\sim}{A}^{-1}$.

On constate immédiatement que si $|\underset{\sim}{A}| = 0$, l'inverse $\underset{\sim}{A}^{-1}$ n'existe pas et on dit que la matrice $\underset{\sim}{A}$ est *singulière* (ou *non inversible*). Par ailleurs, lorsque $\underset{\sim}{A}$ est symétrique, $\underset{\sim}{A}^{-1}$ l'est aussi.

1.3.4 Exemples

On vérifie aisément les déterminants et inverses des *matrices* suivantes :

$$\underset{\sim}{A} = \begin{pmatrix} 2 & 3 \\ 2 & 5 \end{pmatrix} \quad |\underset{\sim}{A}| = 4 \quad \underset{\sim}{A}^{-1} = \begin{pmatrix} 1.25 & -0.75 \\ -0.5 & 0.5 \end{pmatrix}$$

$$\underset{\sim}{A} = \begin{pmatrix} 1.0 & 0.24 & 0.58 \\ 0.24 & 1.0 & 0.73 \\ 0.58 & 0.73 & 1.0 \end{pmatrix} \quad |\underset{\sim}{A}| = 0.2763 \quad \underset{\sim}{A}^{-1} = \begin{pmatrix} 1.690 & 0.664 & -1.465 \\ 0.664 & 2.401 & 2.138 \\ -1.465 & 2.138 & 3.410 \end{pmatrix}$$

$$\underset{\sim}{D} = \mathrm{diag}(4, 5, -2) \quad |\underset{\sim}{D}| = -40 \quad \underset{\sim}{D}^{-1} = \mathrm{diag}\left(\frac{1}{4}, \frac{1}{5}, -\frac{1}{2}\right).$$

1.4 Rang d'une matrice

Le *rang* d'une matrice $\underset{\sim}{A}$ de dimension $r \times c$ est la dimension de la plus grande sous-matrice carrée de déterminant non nul que l'on peut y trouver.

$$0 \leq \text{ Rang } (\underset{\sim}{A}) \leq \min(r, c)$$

Ainsi, le rang de la matrice 3×3

$$\underset{\sim}{A} = \begin{pmatrix} 14 & 3 & 8 \\ 2 & 0 & 2 \\ 2 & 3 & -4 \end{pmatrix}$$

est inférieur ou égal à 3. Toutefois, il ne peut être égal à 3 car $|\underset{\sim}{A}| = 0$. Il est égal à 2 car, par exemple, la sous-matrice carrée $\begin{pmatrix} 14 & 3 \\ 2 & 0 \end{pmatrix}$ de dimension 2×2 est de déterminant non nul.

On peut aussi définir le rang d'une matrice $\underset{\sim}{A}$ de dimension $r \times c$ comme étant le nombre maximum de colonnes ou de lignes de la matrice *linéairement*

indépendantes. On dit que les c vecteurs $\underset{\sim}{\ell}_1, \underset{\sim}{\ell}_2, \ldots, \underset{\sim}{\ell}_c$ sont linéairement indépendants s'il est impossible de trouver des coefficients $k_1, \ldots, k_c$ non tous nuls tels que

$$k_1 \underset{\sim}{\ell}_1 + k_2 \underset{\sim}{\ell}_2 + \ldots + k_c \underset{\sim}{\ell}_c = \underset{\sim}{0}.$$

Dans l'exemple ci-dessus, le rang de $\underset{\sim}{A}$ ne peut être égal à 3 car les 3 colonnes ne sont pas linéairement indépendantes. En effet, en fixant $k_1 = 1, k_2 = -2$ et $k_3 = -1$, on a

$$1 \times \begin{pmatrix} 14 \\ 2 \\ 2 \end{pmatrix} - 2 \times \begin{pmatrix} 3 \\ 0 \\ 3 \end{pmatrix} - 1 \times \begin{pmatrix} 8 \\ 2 \\ -4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Par contre, on peut affirmer que la matrice est de rang 2 car, par exemple, les colonnes 1 et 2 de la matrice sont linéairement indépendantes. En effet, il est impossible de trouver k_1 et k_2 non tous nuls tels que que

$$k_1 \times \begin{pmatrix} 14 \\ 2 \\ 2 \end{pmatrix} + k_2 \times \begin{pmatrix} 3 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Donc, le rang de $\underset{\sim}{A}$ est égal à 2.

Remarques

- Lorsque des vecteurs ne sont pas linéairement indépendants, on dit qu'ils sont *linéairement dépendants*. Il existe entre eux une relation linéaire et l'un peut s'exprimer en fonction des autres.
- La matrice carrée $\underset{\sim}{A}$ de dimension $r \times r$ est singulière si et seulement si elle n'est pas de rang maximum. Donc, si rang $\underset{\sim}{A} < r$, alors $|\underset{\sim}{A}| = 0$ et inversément.
- Le rang d'une matrice carrée donne la véritable dimension de la matrice. Si elle n'est pas de rang maximum, elle contient en quelque sorte des éléments redondants.

1.5 Forme quadratique

Si $\underset{\sim}{A}$ est une matrice symétrique de dimension $r \times r$ et $\underset{\sim}{b}$ un vecteur de dimension r, le triple produit

$$\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b}$$

est un scalaire appelé *forme quadratique*.

On dit que la matrice $\underset{\sim}{A}$ ou la forme quadratique $\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b}$ est *définie positive (dp)* si et seulement si

$$\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b} > 0 \quad \forall \underset{\sim}{b} \neq \underset{\sim}{0}$$

Par contre, si $\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b} \geq 0 \quad \forall \underset{\sim}{b} \neq 0$, la forme est dite *semi-définie positive (sdp)*.

1.6 Valeurs propres et vecteurs propres

1.6.1 Valeurs propres

Soit $\underset{\sim}{A}$ une matrice carrée de dimension $r \times r$ et λ un scalaire. Les *valeurs propres* de la matrice $\underset{\sim}{A}$ sont les solutions de l'équation caractéristique

$$|\underset{\sim}{A} - \lambda \underset{\sim}{I}| = 0$$

c'est-à-dire du polynome d'ordre r

$$|\underset{\sim}{A} - \lambda \underset{\sim}{I}| = b_0 + b_1\lambda + b_2\lambda^2 + \ldots + b_r\lambda^r$$

Dans l'équation ci-dessus, on montre facilement que $b_0 = |\underset{\sim}{A}|$, $b_{r-1} = (-1)^{r-1} tr\underset{\sim}{A}$ et $b_r = (-1)^r$.

On note $\lambda_1, \lambda_2, \ldots, \lambda_r$, les r valeurs propres de la matrice $\underset{\sim}{A}$, solutions de l'équation polynomiale. Ces valeurs propres peuvent être réelles ou complexes. Si on suppose que la matrice $\underset{\sim}{A}$ est symétrique, comme c'est souvent le cas en statistique, alors les valeurs propres sont réelles.

On montre aisément que

$$tr\underset{\sim}{A} = \lambda_1 + \lambda_2 + \ldots + \lambda_r = \sum_{i=1}^{r} \lambda_i$$

$$|\underset{\sim}{A}| = \lambda_1 \times \lambda_2 \times \ldots \times \lambda_r = \prod_{i=1}^{r} \lambda_i$$

Si la matrice $\underset{\sim}{A}$ est *définie positive*, $\lambda_i > 0 \ \forall i$. Si $\underset{\sim}{A}$ est *semi-définie positive*, il existe au moins une valeur propre λ_i nulle. Dans ce dernier cas, la matrice n'est pas invertible puisque $|\underset{\sim}{A}| = 0$. On voit aussi que le rang de $\underset{\sim}{A}$ est le nombre de ses valeurs propres non nulles.

Les valeurs propres (en anglais, "eigen values" ou "latent roots") sont en quelque sorte les atomes constitutifs de la matrice $\underset{\sim}{A}$.

1.6.2 Vecteurs propres

A chaque valeur propre $\lambda_i\,(i = 1, \ldots, r)$ est associé un *vecteur propre* $\underset{\sim}{x}_i$ obtenu en résolvant le système d'équations

$$\underset{\sim}{A}\underset{\sim}{x}_i = \lambda_i \underset{\sim}{x}_i$$

ou encore

$$(\underset{\sim}{A} - \lambda_i \underset{\sim}{I})\underset{\sim}{x}_i = \underset{\sim}{0}$$

La solution $\underset{\sim}{x}_i \neq 0$ existe parce que la matrice $(\underset{\sim}{A} - \lambda_i \underset{\sim}{I})$ est non inversible, son déterminant étant nul . Dans le cas contraire, on aurait obtenu la solution triviale $\underset{\sim}{x}_i = \underset{\sim}{0}$.

La solution $\underset{\sim}{x}_i \neq \underset{\sim}{0}$ n'est toutefois pas unique et il est habituel de normer le vecteur $\underset{\sim}{x}_i$ à l'unité, soit $\|\underset{\sim}{x}_i\| = \underset{\sim}{x}_i^T \cdot \underset{\sim}{x}_i = 1$.

Les vecteurs propres $\underset{\sim}{x}_i$ et $\underset{\sim}{x}_j$ associés à des valeurs propres différentes $\lambda_i \neq \lambda_j$ sont orthogonaux, soit

$$\underset{\sim}{x}_i^T \cdot \underset{\sim}{x}_j = 0 \quad \forall i \neq j$$

Le système des vecteurs propres ("eigen vectors" en anglais) constitue donc un référentiel cartésien.

On montre que $\underset{\sim}{A} = \lambda_1 \underset{\sim}{x}_1 \cdot \underset{\sim}{x}_1^T + \ldots + \lambda_r \underset{\sim}{x}_r \cdot \underset{\sim}{x}_r^T$.

A titre d'exemple, considérons la matrice symétrique de dimension 2×2 suivante :

$$\underset{\sim}{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

Ses valeurs propres sont solutions de l'équation du second degré

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (1-\lambda)(2-\lambda) - 1$$
$$= 1 - 3\lambda + \lambda^2 = 0$$

soient $\lambda_1 = (3 + \sqrt{5})/2 \simeq 2.618$ et $\lambda_2 = (3 - \sqrt{5})/2 \simeq 0.382$. Observons que $\lambda_1 + \lambda_2 = tr\underset{\sim}{A} = 3$ et que $\lambda_1 \times \lambda_2 = |\underset{\sim}{A}| = 1$.

Le vecteur propre correspondant à la valeur propre $\lambda_1 = 2.618$ est obtenu à partir de l'équation

$$\begin{pmatrix} -1.618 & 1 \\ 1 & -0.618 \end{pmatrix} \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

qui a pour solution $x_1 = k$ et $x_2 = 1.618k\,(k \neq 0)$. Pour obtenir une solution unique, on norme le vecteur à l'unité en prenant $k = 0.5257$. D'où $\underset{\sim}{x}_1^T = (0.5257, 0.8507)$.

De même, le vecteur propre associé à la valeur propre $\lambda_2 = 0.382$ est obtenu à partir de l'équation

$$\begin{pmatrix} 0.618 & 1 \\ 1 & 1.618 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

et on trouve $x_1 = k$ et $x_2 = -0.618k\,(k \neq 0)$. Pour obtenir une solution unique, on norme le vecteur à l'unité en prenant $k = 0.8507$. D'où $\underset{\sim}{x}_2^T = (0.8507, -0.5257)$.

On vérifiera que les vecteurs propres $\underset{\sim}{x}_1$ et $\underset{\sim}{x}_2$ sont orthogonaux et que (aux erreurs d'arrondis près)

$$\underset{\sim}{A} = 2.618 \times \begin{pmatrix} 0.5257 \\ 0.8507 \end{pmatrix} (0.5257,\ 0.8507) + 0.382 \times \begin{pmatrix} 0.8507 \\ -0.5257 \end{pmatrix} (0.8507,\ -0.5257)$$

1.7 Equation matricielle

Le calcul matriciel permet l'écriture allégée et la résolution aisée de systèmes d'équations. Ainsi, si $\underset{\sim}{A}$ est une matrice carrée non singulière de dimension $r \times r$, $\underset{\sim}{x}$ et $\underset{\sim}{b}$ des vecteurs de longueur r, on peut aisément résoudre le système d'équation :

$$\underset{\sim}{A} \cdot \underset{\sim}{x} = \underset{\sim}{b}$$

en multipliant les deux membres de l'égalité par l'inverse de la matrice $\underset{\sim}{A}$ et on a

$$\underset{\sim}{x} = \underset{\sim}{A}^{-1} \cdot \underset{\sim}{b}.$$

Ainsi, le système de 3 équations à 3 inconnues

$$\begin{aligned} 8x_1 + 7x_2 - 3x_3 &= 14 \\ 7x_1 - 2x_2 + 4x_3 &= 8 \\ x_1 + 2x_2 - 5x_3 &= -1 \end{aligned}$$

s'écrit sous la forme matricielle

$$\begin{pmatrix} 8 & 7 & -3 \\ 7 & -2 & 4 \\ 1 & 2 & -5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 14 \\ 8 \\ -1 \end{pmatrix}.$$

L'inverse de la matrice $\underset{\sim}{A}$ s'obtient comme décrit précédemment. Le déterminant de $\underset{\sim}{A}$ vaut $|\underset{\sim}{A}| = 241$ et la matrice des cofacteurs s'écrit

$$\begin{pmatrix} 2 & 39 & 16 \\ 29 & -37 & -9 \\ 22 & -53 & -65 \end{pmatrix}.$$

La matrice inverse vaut donc (après transposition de la matrice des cofacteurs)

$$\underset{\sim}{A}^{-1} = \frac{1}{241}\begin{pmatrix} 2 & 29 & 22 \\ 39 & -37 & -53 \\ 16 & -9 & -65 \end{pmatrix} = \begin{pmatrix} 0.00830 & 0.120 & 0.0913 \\ 0.162 & -0.154 & -0.220 \\ 0.0664 & -0.0373 & -0.270 \end{pmatrix}.$$

Le vecteur $\underset{\sim}{x}$ solution de l'équation de départ est donc obtenu à partir de l'équation

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.00830 & 0.120 & 0.0913 \\ 0.162 & -0.154 & -0.220 \\ 0.0664 & -0.0373 & -0.270 \end{pmatrix}\begin{pmatrix} 14 \\ 8 \\ -1 \end{pmatrix}.$$

Dès lors, $x_1 = 0.00830 \times 14 + 0.120 \times 8 + 0.0913 \times (-1) = 0.985$. De même, $x_2 = 1.256$ et $x_3 = 0.901$.

En conclusion, $\underset{\sim}{x}^T = (0.985, 1.256, 0.901)$.

Chapitre 2

Matrice d'observations

2.1 Introduction

En statistique multivariée, on dispose au départ d'une matrice d'observations dont les lignes correspondent aux sujets analysés et les colonnes aux variables envisagées. Avant de définir formellement cette matrice, il est bon de rappeler certaines notions de base utiles pour la suite. Ainsi, on définira les concepts de population, d'échantillon et de variables. On envisagera les différents types de variables et les transformations qu'il convient parfois de leur appliquer avant de construire la matrice d'observations. On terminera le chapitre par quelques représentations graphiques des données multivariées.

2.2 Population et échantillon

Dans tout problème statistique, il convient de définir clairement la population à laquelle on s'intéresse et de laquelle on tire un échantillon.

2.2.1 Population

Une population est un ensemble d'individus (ou d'objets) qui possèdent au moins une caractéristique en commun. On désigne par N l'effectif de la population. Celui-ci est souvent assimilé à l'infini ($N \approx \infty$), de sorte qu'il n'apparaît pas dans les formules mathématiques. A titre d'exemples, citons la population des étudiants de l'université de Liège, celle des villes de Belgique, ou encore celle des patients du CHU.

On ne peut réellement commencer une analyse statistique tant que la population concernée n'a pas été clairement définie. Notons que l'élément (sujet ou objet) de la population sur lequel vont porter les observations ou

les mesures est appelé *unité statistique*. Dans les quelques exemples cités, il s'agit respectivement de l'étudiant, de la ville et du patient.

2.2.2 Echantillon

En pratique, comme la population est infinie, on ne peut l'étudier complètement. On se résout dès lors à tirer un échantillon, c'est-à-dire un sous-ensemble de la population concernée. On désigne par n l'effectif de l'échantillon ; ce nombre est par définition fini. Il n'est pas facile d'extraire un échantillon d'une population. Il faut avoir recours à des méthodes d'échantillonnage rigoureuses, sans quoi l'échantillon pourrait ne pas être représentatif de la population, entraînant de ce fait des conclusions biaisées sur la population.

2.2.3 Echantillonnage

L'échantillonnage est un mécanisme (processus) aléatoire permettant de tirer des échantillons d'une population. Il existe plusieurs méthodes d'échantillonnage dont la plus connue est l'échantillonnage simplement fortuit. Celui-ci répond aux trois conditions suivantes :

1. l'effectif n est fixé à l'avance ;
2. les tirages successifs se font au hasard (chaque élément de la population a la même chance d'être tiré) et de façon indépendante (le tirage d'un élément ne peut conditionner le suivant) ;
3. les tirages successifs se font d'une population invariante (les tirages successifs ne peuvent modifier la population).

Comme les populations sont infinies, la condition (3) est toujours remplie. Dans le cas contraire, il convient de remettre l'élément tiré dans la population ; ceci donne l'apparence d'une population qui ne bouge pas mais signifie aussi qu'un même élément peut être extrait plusieurs fois. Il existe d'autres types d'échantillonnage, comme l'échantillonnage stratifié, l'échantillonnage systématique ou l'échantillonnage en grappes. Dans tous les cas, on a recours à l'ordinateur pour générer aléatoirement les éléments à extraire de la population.

2.3 Variables

2.3.1 Définition

En statistique, les variables sont les caractéristiques (critères, facteurs) des éléments de la population auxquels on s'intéresse. Dans tout problème

statistique, celles-ci doivent être définies avec précision et soin. A titre d'exemple, pour la population des étudiants de l'université de Liège, on peut s'intéresser à l'âge, au sexe, à la faculté à laquelle ils appartiennent, à la race, ou encore au grade académique obtenu. Pour les villes de Belgique, les variables pourraient être le nombre d'habitants, le niveau socio-économique, le taux de pollution annuel, la fréquence d'affections respiratoires. Enfin, pour les patients du CHU, on aurait l'âge, le sexe, le type de pathologie, la durée et le coût d'hospitalisation, l'issue, la destination à la sortie de l'hôpital.

Les variables peuvent être de différents types. Il convient notamment de distinguer les variables quantitatives des variables qualitatives car les secondes requièrent souvent une transformation avant d'être utilisées dans une analyse statistique multivariée.

2.3.2 Variable quantitative

Une variable quantitative exprime une quantité. Les valeurs prises par une variable quantitative résultent le plus souvent d'une mesure (avec un appareil) ou d'un comptage ; elles sont donc numériques et peuvent être reportées telles quelles dans la matrice d'observations.

Lorsque la variable quantitative concerne un comptage, on parle de variable "discrète" car elle prend un nombre fini de valeurs distinctes. C'est le cas du nombre d'échecs pour un étudiant, du nombre d'agents de police dans une ville, du nombre de médicaments pris par un patient.

Lorsque la variable quantitative concerne une mesure, on parle de variable "continue" car celle-ci peut en principe prendre toutes les valeurs possibles dans un intervalle donné, donc une infinité. Seule la précision de l'appareil de mesure donne l'impression que la variable n'est pas continue. Citons à titre d'exemple, la taille (cm) d'un étudiant, la concentration d'ozone (mg/mm^3) dans l'air d'une ville, le taux de cholestérol (g/l) d'un patient. On notera enfin que les variables quantitatives continues sont en général affectées d'unités de mesure (cm, kg, mmHg, mg/mm^3, etc).

2.3.3 Variable binaire

Une variable binaire est une variable qui ne prend que deux états possibles, par exemple absent ou présent, échec ou réussite, homme ou femme, sain ou malade. Il est classique de coder une variable binaire par les nombres 0 et 1. Ainsi, 0 = absence et 1 = présence, 0 = échec et 1 = réussite, 0 = homme et 1 = femme, 0 = sain et 1 = malade. La solution d'un problème statistique ne dépend pas de la manière dont on code les deux états mais il faut se rappeler ce qui a été codé 0 et ce qui a été codé 1. Les variables

binaires sont assimilées à des variables quantitatives (numériques) dans la mesure où l'on peut faire des calculs sur ses valeurs (0 et 1).

2.3.4 Variable qualitative

Une variable qualitative traduit une qualité. Elle ne se mesure pas mais en général elle s'observe. Les valeurs prises par une variable qualitative, appelées classes, catégories ou modalités, sont des noms et non des nombres ! C'est la raison pour laquelle on parle aussi de variables nominales. A titre d'exemple, citons la race d'un étudiant (blanc, noir, jaune), le type de ville (petite, moyenne ou grande), l'état civil d'un patient (célibataire, marié, divorcé, veuf). Lorsque les modalités de la variable qualitative respectent un ordre, on dit que la variable est ordinale. C'est le cas du type de ville (petite, moyenne, grande) ou du résultat à l'examen d'un étudiant (ajournement, satisfaction, distinction, grande distinction, plus grande distinction). En général, les modalités des variables qualitatives sont mutuellement exclusives, c'est-à-dire qu'on ne peut avoir deux modalités différentes à la fois. Ainsi, on ne peut être à la fois noir et jaune, c'est l'un ou l'autre. Il arrive que pour certaines variables qualitatives, les modalités ne soient pas disjointes. C'est souvent le cas des variables que l'on rencontre dans les questionnaires. Par exemple, la variable qualitative "hobby" peut prendre les valeurs suivantes : sport, lecture, théâtre, cinéma, jardinage, musique. Un individu peut dès lors choisir une ou plusieurs modalités. Observons enfin qu'une variable binaire est une variable qualitative à deux modalités. Comme deux modalités sont toujours ordonnées, une variable binaire est donc aussi ordinale.

2.3.5 Variable catégorisée

Il arrive parfois qu'une variable continue ne soit ou ne puisse être mesurée telle quelle. C'est le cas notamment lorsque la variable présente un caractère confidentiel ou personnel comme le salaire mensuel d'une personne ou le nombre de patients que voit un médecin par semaine. On divise alors l'espace des valeurs de la variable en catégories et on choisit la catégorie qui convient comme réponse. On a alors affaire à une variable catégorisée, qui peut être considérée comme une variable qualitative ordinale. Par exemple, la variable "salaire mensuel" pourrait être catégorisée comme suit : moins de 500 €/mois, de 500 à 1000 €/mois, de 1000 à 2000 €/mois, de 2000 à 3000 €/mois, plus de 3000 €/mois. La patientèle d'un médecin aurait pour catégories : moins de 30 patients/semaine, 30-60 patients/semaine, 60-90 patients/semaine et $>$ 90 patients/semaine. La catégorisation d'une variable quantitative conduit nécessairement à une perte d'informations.

2.4 Données ou observations

Il ne faut pas confondre variables et données. Les variables sont des concepts théoriques, les données sont des observations réelles. Une donnée est la mesure ou le comptage d'une variable quantitative ou l'observation d'une variable binaire ou qualitative. Ainsi, si la variable étudiée est le poids (kg), une donnée correspond à la mesure du poids chez un individu particulier (par exemple, 73 kg). De même, si la variable étudiée est le grade à l'examen, alors le grade "distinction" obtenu par un étudiant est une donnée ou une observation.

En statistique multivariée, toutes les données doivent être numériques, c'est-à-dire des nombres sur lesquels on va pouvoir faire des calculs (addition, soustraction, multiplication, division). Les variables quantitatives ne posent évidemment pas de problèmes, puisque les observations qui en résultent sont toujours des nombres. Les variables binaires sont également utilisables car les modalités sont codées 0 et 1, deux valeurs numériques.

Il en va différemment des variables qualitatives. Certes, les modalités peuvent être numérotées (par exemple, 1 = célibataire, 2 = marié, 3 = divorcé, 4 = veuf) mais il apparaît immédiatement qu'on ne peut faire des opérations arithmétiques sur ces nombres. Par ailleurs, rien n'empêche d'utiliser une autre numérotation (1 = veuf, 2 = divorcé, 3 = célibataire et 4 = marié). Lorsque la variable qualitative est ordinale, l'ordre des modalités impose la numérotation (par exemple, 1 = petite, 2 = moyenne, 3 = grande). Dans ce cas, il n'est pas tout à fait erroné de faire des calculs sur les valeurs obtenues. C'est le cas de toutes les échelles d'évaluation à 3 ou plusieurs points.

Pour traiter les variables qualitatives de façon numérique, il faut les remplacer par des variables binaires. On procède comme suit selon que les modalités sont mutuellement exclusives ou non. Lorsque les modalités ne sont pas disjointes (par exemple, la variable "hobby"), on associe une variable binaire à chaque modalité. Donc, on définit la variable binaire "sport" (0 = non, 1 = oui), la variable binaire "lecture" (0 = non, 1 = oui) et ainsi de suite pour les autres modalités. Il y a donc autant de variables binaires qu'il y a de modalités. Par contre, lorsque les modalités sont disjointes, on associe une variable binaire à la première modalité, une variable binaire à la seconde modalité et ainsi de suite jusqu'à l'avant-dernière. On n'associe donc pas de variable binaire à la dernière modalité, considérée comme la modalité de référence. La solution du problème ne dépend pas de la modalité qui est prise comme référence. Au total, on a donc autant de variables binaires qu'il y a de modalités moins une. Ainsi, à titre d'exemple, on définit une variable binaire pour la race blanche (0 = non, 1 = oui) et une variable binaire pour

la race noire (0 = non, 1 = oui). La race jaune est prise comme référence. Dans ce cas, si un sujet est de race jaune, il aura la valeur 0 pour chacune des modalités blanche et noire. En clair, il suffit de connaître deux modalités et la troisième est automatiquement connue.

Le processus de remplacement d'une variable qualitative par des variables binaires s'appelle la booléanisation car les variables binaires sont aussi souvent appelées *variables booléennes* (algèbre binaire $0-1$ de Boole). Par ce procédé, plus aucune variable ne prend des valeurs nominales mais au contraire des valeurs numériques.

2.5 Matrice d'observation

Dans la suite, on désigne par $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ le vecteur des p variables "numériques" du problème multivarié. Ceci présuppose que les variables qualitatives ont été transformées en variables binaires correspondantes.

2.5.1 Définition

La *matrice d'observations* (ou matrice des données) est la matrice obtenue par l'observation (ou la mesure) du vecteur $p-$varié $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ chez un échantillon de sujets ou d'objets d'effectif n extrait de la population étudiée.

La matrice d'observations possède donc n lignes (les sujets ou objets) et p colonnes (les variables). Il s'agit d'une matrice $n \times p$ que l'on écrit sous la forme

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}. \tag{2.1}$$

Dans cette matrice, l'élément x_{ij} représente l'observation chez le sujet (ou l'objet) i de la variable $X_j (i = 1, \ldots, n; j = 1, \ldots, p)$. Il s'agit d'une matrice rectangulaire car, en général, il y a plus de sujets que de variables ($n > p$). Toutes les valeurs de la matrice (2.1) sont des nombres réels.

Les dimensions de la matrice, respectivement n et p, sont des éléments essentiels du problème multivarié.

La matrice $\underset{\sim}{X}_{n\times p}$ est complète si aucun des éléments x_{ij} qui la constituent n'est manquant ; dans le cas contraire, elle est dite *incomplète* et pose problème pour la suite des calculs.

2.5.2 Exemples

Considérons un échantillon de 10 étudiants extrait de la population des étudiants de l'Université de Liège et envisageons les variables âge, sexe, race et grade en fin d'année académique. Comme indiqué précédemment, la variable race doit être remplacée par 2 variables binaires "Blanc" et "Noir". Pour la variable ordinale "grade à l'examen", on peut numéroter les modalités car celles-ci sont ordonnées. Les variables âge (années) et sexe (0=homme, 1=femme) ne posent pas de problème. Voici la matrice d'observations :

$$\underset{\sim}{X}_{10\times 5} = \begin{pmatrix} 18 & 0 & 1 & 0 & 1 \\ 19 & 1 & 1 & 0 & 1 \\ 24 & 1 & 1 & 0 & 5 \\ 21 & 0 & 0 & 0 & 3 \\ 22 & 0 & 0 & 1 & 2 \\ 26 & 0 & 1 & 0 & 3 \\ 19 & 1 & 1 & 0 & 2 \\ 17 & 1 & 1 & 0 & 2 \\ 20 & 1 & 0 & 0 & 2 \\ 20 & 1 & 1 & 0 & 4 \end{pmatrix}$$

dans laquelle les étudiants correspondent aux lignes et dont les variables (colonnes) sont X_1 = âge (années), X_2 = sexe (0=masculin, 1=féminin), X_3 = race blanche (0=non, 1=oui), X_4 = race noire (0=non, 1=oui), X_5 = grade (1 = A, 2 = S ; 3 = D, 4 = GD, 5 = PGD). Notons que le sujet N°2 a 19 ans, il est de sexe féminin, de race blanche et a été ajourné. Par ailleurs, le sujet N°4, âgé de 21 ans, est de sexe masculin, de race jaune et a obtenu une distinction.

En annexe, on trouvera d'autres exemples de matrices d'observations. Ainsi, le célèbre statisticien R.A. Fisher, qui s'est intéressé à la population des Iris Setosa, a mesuré quatre variables sur un échantillon de 50 iris. Les variables étaient la longueur des sépales (X_1), la largeur des sépales (X_2), la longueur des pétales (X_3) et la largeur des pétales (X_4). La matrice des observations est donc de dimension 50×4, puisque $n = 50$ et $p = 4$.

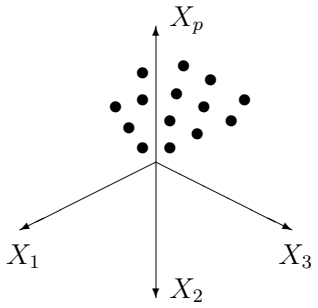
2.5.3 Nuage de points

En associant à chaque variable $X_1, \ldots X_p$ d'un problème multivarié un axe orienté, on définit dans l'espace euclidien à p dimensions $\mathbb{R}^p$ un référentiel d'axes orthogonaux. L'observation multivariée du vecteur $\underset{\sim}{X}$ chez le sujet i

$$\underset{\sim}{x}_i^T = \begin{pmatrix} x_{i1}, & \ldots, & x_{ip} \end{pmatrix}. \tag{2.2}$$

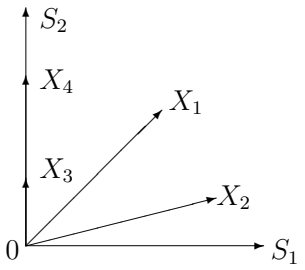
c'est-à-dire la i-ème ligne de la matrice d'observations $\underset{\sim}{X}_{n\times p}$, peut dès lors être considérée comme un point de $\mathbb{R}^p$, appelé espace des observations. En d'autres termes, à toute matrice d'observations $\underset{\sim}{X}_{n\times p}$ correspond un nuage de n points dans l'espace à p dimensions $\mathbb{R}^p$. Il y a autant de points dans le nuage qu'il y a de lignes dans la matrice d'observations.

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \Leftrightarrow$$



Il suffit pour s'en convaincre de l'appliquer au cas classique de 2 variables et du nuage de points dans un graphique à 2 dimensions. Il en est de même pour 3 variables. Au-delà de 3 variables, une représentation classique n'est plus possible mais on peut l'imaginer. A l'inverse, on pourrait associer un axe à chaque sujet ($i = 1, \dots, n$) et travailler dans l'espace $\mathbb{R}^n$. L'observation de la variable X_j pour chaque sujet, à savoir le j-ème vecteur colonne $(x_{1j}, x_{2j}, \dots, x_{nj})^T$ correspond à un point dans l'espace $\mathbb{R}^n$. En conséquence, il y a p points dans l'espace $\mathbb{R}^n$, un pour chaque variable du problème multivarié. Cette représentation est plus difficile à imaginer en pratique. A titre d'exemple, l'observation de quatre variables chez deux sujets (S_1 et S_2) donne lieu à la figure suivante :

$$\underset{\sim}{X}_{2\times 4} = \begin{pmatrix} 3 & 5 & 0 & 0 \\ 4 & 1 & 2 & 6 \end{pmatrix} \quad \Leftrightarrow$$



En conclusion, une matrice d'observations $\underset{\sim}{X}_{n\times p}$ correspond à un nuage de n points dans l'espace $\mathbb{R}^p$ ou de p points dans l'espace $\mathbb{R}^n$. Dans $\mathbb{R}^p$, les points voisins ont un profil semblable, tandis que dans $\mathbb{R}^n$ des variables situées dans des directions voisines sont corrélées.

2.6 Représentations graphiques

Même si toute analyse statistique multivariée doit débuter par une analyse statistique univariée de chaque variable, les statisticiens ont toujours cherché à représenter graphiquement les observations multivariées.

A titre d'exemple, considérons 6 étudiants de 1e année de médecine et leurs cotes d'examens (0-20) pour les matières suivantes : X_1=Chimie (Ch), X_2=Physique (Ph), X_3=Sciences biomédicales (Sb), X_4=Santé et société (Ss) et X_5=Techniques d'apprentissage (Ta). Les données sont reprises dans le tableau ci-dessous :

Etudiant	Ch	Ph	Sb	Ss	Ta
1	19	19	15	14	16
2	15	13	12	14	12
3	12	10	11	15	11
4	5	5	12	14	14
5	4	3	8	11	9
6	1	0	2	9	10

Il y a différentes manières de représenter le profil de chaque étudiant. Il est utile au préalable de ramener toutes les observations dans l'intervalle [0-1] par la transformation

$$Y = \frac{X - min}{max - min}$$

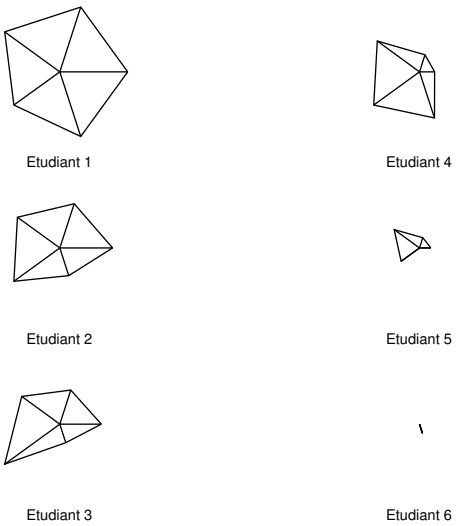
où min et max représentent respectivement la cote la plus basse et la plus élevée dans chaque matière.

2.6.1 Les glyphes

Un *glyphe* est un cercle fixe avec p rayons extérieurs au cercle de longueur variable correspondant à la valeur de chaque variable. On dessine un glyphe pour chacun des n sujets.

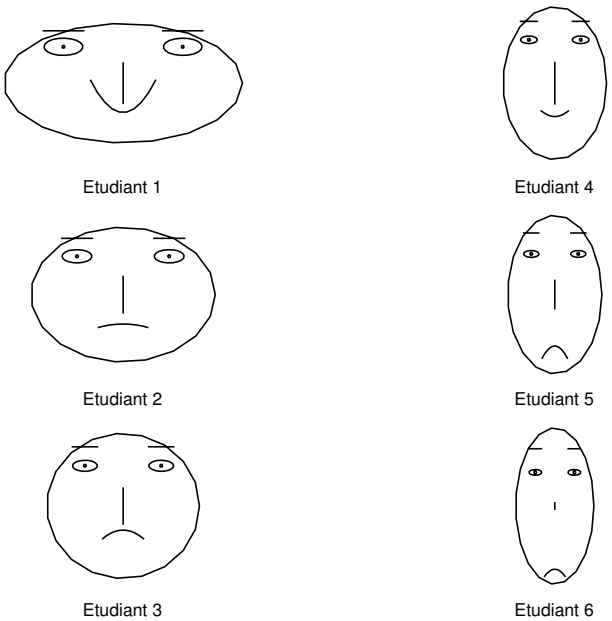
2.6.2 Les étoiles

Une *étoile* est semblable à un glyphe sauf que les rayons sont de longueur égale et inscrits dans le cercle. Les valeurs des variables sur les différents rayons sont joints entre eux pour former un polygone qui ressemble à une étoile.



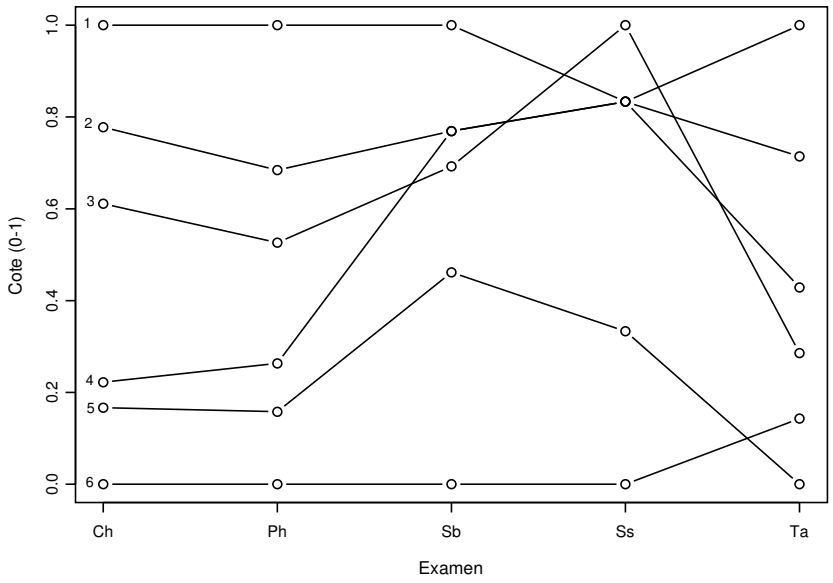
2.6.3 Les faces de Chernoff

La méthode des *faces de Chernoff* consiste à associer à chaque variable une caractéristique différente du visage humain comme la longueur du nez, la forme du visage, la taille des yeux, celle des oreilles et ainsi de suite. Malheureusement, avec cette technique, les faces changent en interchangeant les variables et il faut parfois essayer plusieurs approches avant d'avoir la bonne.



2.6.4 Les profils

Cette méthode est similaire à celle des étoiles sauf que les variables sont positionnées séquentiellement sur l'axe des abscisses. On joint entre elles les valeurs de chaque variable et les segments de droite ainsi obtenus forment un *profil*. Dans le cas de variables dont les données de valeurs sont très différentes, cette approche requiert une certaine standardisation.

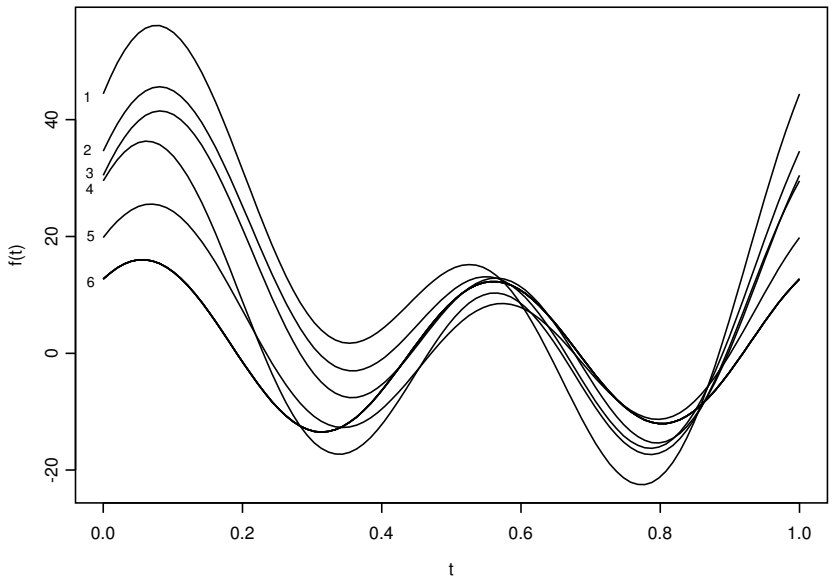


2.6.5 Les graphiques de Fourier

Cette méthode due à Andrews consiste à représenter une observation multivariée $\underset{\sim}{x}_i^T = (x_{i1}, \ldots, x_{ip})$ par la série finie de Fourier

$$f_i(t) = x_{i1}/\sqrt{2} + x_{i2} \sin t + x_{i3} \cos t + x_{i4} \sin 2t + x_{i5} \cos 2t + \ldots$$

et de reporter le graphe $f_i(t)$ dans l'intervalle $-\pi < t < \pi$ (ou encore en remplaçant t par $2\pi t$ dans l'intervalle $0 < t < 1$).



2.7 Observation multivariées aberrantes

Comme en statistique univariée, une observation peut être aberrante au niveau multivarié, c'est-à-dire provenir d'une autre population que celle qu'on étudie. Dans ces circonstances, l'observation se démarque nettement des autres observations. Elle présente un profil atypique.

De telles observations peuvent affecter sérieusement les résultats de l'analyse statistique. Il est donc important de pouvoir les déceler, les corriger ou éventuellement les éliminer de la matrice d'observations. Les représentations graphiques décrites à la section 2.6. permettent parfois de mettre en évidence des observations aberrantes ou extrêmes. On exposera dans la suite d'autres approches pour aborder ce problème.

Chapitre 3

Moyenne et dispersion

3.1 Introduction

Dans ce chapitre, on se propose de résumer une matrice d'observations par quelques paramètres caractéristiques. C'est la première étape de toute analyse statistique multivariée. R.A. Fisher disait que la statistique est la discipline qui étudie en premier lieu les méthodes de réduction de données. On introduit la notion de vecteur moyen qui étend au niveau multidimensionnel le concept de moyenne arithmétique. On définit ensuite la matrice de dispersion qui caractérise la variabilité d'un échantillon multivarié et qui généralise la notion de variance. En pratique, la matrice de dispersion n'est pas facile à interpréter ; c'est la raison pour laquelle on introduit la matrice de corrélations. On termine le chapitre avec la définition de la distance généralisée de Mahalanobis qui permet de calculer la distance entre une observation multivariée et le point moyen, en tenant compte des associations entre les différentes variables. Cette distance généralise la notion de distance réduite univariée.

3.2 Vecteur moyen

Soit la matrice d'observations $\underset{\sim n\times p}{X}$ définie au Chapitre 2. Les lignes de cette matrice constituent les observations p-variées des n sujets. On peut donc écrire la matrice d'observations sous la forme

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{i1} & \dots & x_{ip} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{x}_1^T \\ \dots \\ \underset{\sim}{x}_i^T \\ \dots \\ \underset{\sim}{x}_n^T \end{pmatrix} \tag{3.1}$$

où $\underset{\sim}{x}_i^T = (x_{i1}, \dots, x_{ip})$ est l'observation du vecteur $\underset{\sim}{X}^T = (X_1, \dots, X_p)$ chez le sujet i.

Pour rappel, si l'on s'intéresse uniquement aux observations de la variable X_j, c'est-à-dire la colonne j de la matrice (3.1), la moyenne de ces observations s'écrit

$$\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj} \quad (j = 1, \dots, p) \tag{3.2}$$

ce qui revient à sommer les observations de la colonne j et à diviser par l'effectif n.

Par définition, le vecteur moyen de l'échantillon multivarié est le vecteur des moyennes dans p variables. On a donc

$$\underset{\sim}{\bar{x}}^T = (\bar{x}_1, \dots, \bar{x}_p) \tag{3.3}$$

En calcul matriciel, l'expression (3.3) peut aussi s'écrire

$$\underset{\sim}{\bar{x}} = \frac{1}{n}\sum_{k=1}^{n} \underset{\sim}{x}_k \tag{3.4}$$

par analogie avec la formule univariée (3.2). On somme les n vecteurs d'observations et on divise par n.

Dans l'espace à p dimensions $\mathbb{R}^p$, le point moyen n'est autre que le centre de gravité (au sens physique du terme) du nuage de points. En effet, $\sum_k(\underset{\sim}{x}_k - \underset{\sim}{\bar{x}}) = \underset{\sim}{0}$. Il s'agit d'un paramètre de position car il situe le nuage de points dans l'espace $\mathbb{R}^p$.

3.3 Matrice de dispersion

Le vecteur moyen ne fournit aucune information sur la dispersion du nuage de points dans l'espace à p dimensions. Il est nécessaire à cet effet d'introduire la notion de matrice de variances-covariances ou matrice de dispersion.

Pour rappel, lorsqu'on s'intéresse à deux variables X_i et X_j, on définit la covariance qui mesure l'association entre elles. La covariance $cov(X_i, X_j)$ s'écrit

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{3.5}$$

ou encore

$$s_{ij} = \frac{1}{n-1} \left\{ \sum_{k=1}^{n} x_{ki} x_{kj} - \frac{\left(\sum_{k=1}^{n} x_{ki}\right)\left(\sum_{k=1}^{n} x_{kj}\right)}{n} \right\}. \tag{3.6}$$

Une covariance positive (négative) est le signe d'une relation linéaire croissante (décroissante) entre les deux variables, tandis qu'une covariance voisine de 0 est le signe d'une absence de relation.

Lorsqu'on calcule la covariance d'une variable X_i avec elle-même, soit $cov(X_i, X_i)$, les formules (3.5) et (3.6) se simplifient et deviennent respectivement :

$$s_{ii} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2 \tag{3.7}$$

et

$$s_{ii} = \frac{1}{n-1} \left\{ \sum_{k=1}^{n} x_{ki}^2 - \frac{\left(\sum_{k=1}^{n} x_{ki}\right)^2}{n} \right\}. \tag{3.8}$$

On retrouve ainsi la variance $s_i^2 = var(X_i)$ de la variable X_i. On constate donc que la variance est un cas particulier de la covariance. Pour rappel, une variance n'est jamais négative.

Par définition, la matrice de dispersion $\underset{\sim}{S}$ est la matrice carrée de dimension $p \times p$ qui contient sur la diagonale les variances des p variables et en-dehors de la diagonale les covariances. Elle est dès lors appelée *matrice de variances-covariances*. On a

$$\underset{\sim}{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}. \tag{3.9}$$

Puisque $s_{ij} = s_{ji}$, la matrice $\underset{\sim}{S}$ est symétrique.

En calcul matriciel, l'expression (3.9) peut aussi s'écrire

$$\underset{\sim}{S} = \frac{1}{n-1} \sum_{k=1}^{n} (\underset{\sim}{x}_k - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_k - \underset{\sim}{\bar{x}})^T \tag{3.10}$$

ou encore

$$\underset{\sim}{S} = \frac{1}{n-1}\left\{\sum_{k=1}^{n} \underset{\sim}{x}_k \underset{\sim}{x}_k^T - \frac{\left(\sum_{k=1}^{n} \underset{\sim}{x}_k\right)\left(\sum_{k=1}^{n} \underset{\sim}{x}_k\right)^T}{n}\right\} \tag{3.11}$$

par analogie avec les formules univariées (3.5) et (3.6).

La matrice de dispersion caractérise la forme du nuage de points dans $\mathbb{R}^p$ autour de la moyenne $\underset{\sim}{\bar{x}}$.

Pour calculer la matrice de dispersion, il suffit de faire la somme des carrés des éléments de chaque colonne de la matrice d'observations et la somme des produits des éléments des colonnes prises 2 à 2. On définit ainsi une matrice $\underset{\sim}{A}$ appelée *matrice des sommes de carrés et produits croisés corrigés*, dont l'élément générique s'écrit

$$a_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj} - \frac{\left(\sum_{k=1}^{n} x_{ki}\right)\left(\sum_{k=1}^{n} x_{kj}\right)}{n} \qquad (i,j = 1,\ldots,p) \tag{3.12}$$

En vertu de la relation (3.6), on a

$$s_{ij} = \frac{1}{n-1}\, a_{ij}$$

et donc

$$\underset{\sim}{S} = \frac{1}{n-1}\, \underset{\sim}{A} \tag{3.13}$$

Afin de simplifier les notations, posons par convention

$$\sum x_j = \sum_{k=1}^{n} x_{kj} \qquad (j = 1,\ldots,p)$$

$$\sum x_j^2 = \sum_{k=1}^{n} x_{kj}^2 \qquad (j = 1,\ldots,p)$$

$$\sum x_i x_j = \sum_{k=1}^{n} x_{ki}x_{kj} \qquad (i,j = 1,\ldots,p; i < j)$$

respectivement la somme des éléments de la colonne j, la somme des carrés des éléments de la colonne j et la somme des produits des éléments des colonnes i et j.

Dans ces circonstances,

$$\begin{aligned} \bar{x}_i &= \sum x_i/n \\ a_{ij} &= \sum x_i x_j - \frac{\left(\sum x_i\right)\left(\sum x_j\right)}{n} \quad (i,j = 1,\ldots,p) \\ s_{ij} &= a_{ij}/(n-1) \end{aligned} \tag{3.14}$$

3.4 Matrice de corrélations

Même si la matrice de variances-covariances $\underset{\sim}{S}$ revêt un grand intérêt théorique, son interprétation en pratique s'avère difficile. En effet, les covariances sont des grandeurs négatives ou positives pourvues d'unités (le produit des unités des deux variables considérées). Il en va de même des variances. C'est la raison pour laquelle on introduit la notion de corrélation entre deux variables X_i et X_j, soit $\text{corr}(X_i, X_j)$, définie comme suit :

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} \cdot s_{jj}}} \qquad (i, j = 1, \ldots, p) \tag{3.15}$$

ou encore, en utilisant les notations condensées (3.14)

$$r_{ij} = \frac{\sum x_i x_j - \frac{\left(\sum x_i\right)\left(\sum x_j\right)}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]\left[\sum x_j^2 - \frac{\left(\sum x_j\right)^2}{n}\right]}} \tag{3.16}$$

Pour calculer la corrélation entre deux variables, cinq quantités sont nécessaires, à savoir $\sum x_i, \sum x_j, \sum x_i^2, \sum x_j^2$ et $\sum x_i x_j$. Notons enfin que si $i = j, r_{ii} = 1$ $(i = 1, \ldots, p)$.

Par définition, la *matrice de corrélations* $\underset{\sim}{R}$ est la matrice carrée de dimension $p \times p$ qui contient toutes les corrélations des variables prises 2 à 2. On a

$$\underset{\sim}{R} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix} \tag{3.17}$$

Sur la diagonale principale, on retrouve des valeurs égales à 1 et en dehors de la diagonale les corrélations. Puisque $r_{ij} = r_{ji}$, la matrice est symétrique.

En calcul matriciel, on peut montrer que

$$\underset{\sim}{R} = \underset{\sim}{D}^{-1} \, \underset{\sim}{S} \, \underset{\sim}{D}^{-1} \tag{3.18}$$

où $\underset{\sim}{D} = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$ est la matrice diagonale des écarts-types des variables, puisque $\sqrt{s_{ii}} = \sqrt{s_i^2} = s_i$ est l'écart-type de la variable X_i. En conclusion, si on connaît $\underset{\sim}{S}$, on peut calculer $\underset{\sim}{R}$ mais l'inverse n'est possible que si on connaît la matrice $\underset{\sim}{D}$.

Pour rappel, une corrélation est un nombre pur toujours compris entre –1 et +1. Une corrélation positive (négative) indique une relation linéaire croissante (décroissante) entre deux variables, tandis qu'une corrélation nulle traduit l'absence de relation.

3.5 Exemples

Reprenons l'exemple de la matrice d'observations 8×3 donnée à la section 1.1.3. Il s'agit de l'observation chez 8 sujets de 3 variables, X_1 = taille (cm), X_2 = poids (kg) et X_3 = circonférence du bras (cm).

$$\underset{\sim}{X}_{8\times 3} = \begin{pmatrix} 167.9 & 71.8 & 30.0 \\ 183.8 & 75.1 & 29.4 \\ 172.9 & 58.0 & 26.0 \\ 175.5 & 58.4 & 25.7 \\ 176.4 & 67.7 & 27.9 \\ 168.5 & 75.2 & 31.7 \\ 178.0 & 67.3 & 27.4 \\ 178.0 & 71.3 & 29.0 \end{pmatrix}.$$

Dans l'espace à 3 dimensions, cette matrice représente un nuage de 8 points. Calculons le vecteur moyen $\underset{\sim}{\bar{x}}$ ainsi que les matrices de dispersion $\underset{\sim}{S}$ et de corrélations $\underset{\sim}{R}$.

Dans ce problème $n = 8$ et $p = 3$.

A cet effet, calculons d'abord les sommes, sommes de carrés et sommes de produits croisés des observations, soit au total $p + p + p(p-1)/2 = 9$ quantités.

$$\begin{aligned} \textstyle\sum x_1 &= 167.9 + 183.8 + \ldots + 178.0 = 1401.0 \\ \textstyle\sum x_2 &= 71.8 + 75.1 + \ldots + 71.3 = 544.8 \\ \textstyle\sum x_3 &= 30.0 + 29.4 + \ldots + 29.0 = 227.1 \end{aligned}$$

$$\begin{aligned} \textstyle\sum x_1^2 &= (167.9)^2 + (183.8)^2 + \ldots + (178.0)^2 = 245544.7 \\ \textstyle\sum x_2^2 &= (71.8)^2 + (75.1)^2 + \ldots + (71.3)^2 = 37421.12 \\ \textstyle\sum x_3^2 &= (30.0)^2 + (29.4)^2 + \ldots + (29.0)^2 = 6475.91 \end{aligned}$$

$$\begin{aligned} \textstyle\sum x_1x_2 &= (167.9 \times 71.8) + \ldots + (178.0 \times 71.3) = 95420.28 \\ \textstyle\sum x_1x_3 &= (167.9 \times 30.0) + \ldots + (178.0 \times 29.0) = 39748.68 \\ \textstyle\sum x_2x_3 &= (71.8 \times 30.0) + \ldots + (71.3 \times 29.0) = 15555.21 \end{aligned}$$

Le vecteur moyen s'écrit immédiatement

$$\begin{aligned} \underset{\sim}{\bar{x}}^T &= \left(\tfrac{1401}{8}, \tfrac{544.8}{8}, \tfrac{227.1}{8}\right) \\ &= (175.1, 68.1, 28.4) \end{aligned}$$

Pour calculer la matrice de dispersion, déterminons d'abord la matrice $\underset{\sim}{A}$ de dimension 3×3 en utilisant la formule (3.12)

$$a_{11} = \sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} = 245544.7 - \frac{(1401)^2}{8} = 194.60$$

$$a_{12} = \sum x_1 x_2 - \frac{(\sum x_1)(\sum x_2)}{n} = 95420.28 - \frac{1401 \times 544.8}{8} = 12.18$$

$$a_{13} = \sum x_1 x_3 - \frac{(\sum x_1)(\sum x_3)}{n} = 39748.68 - \frac{1401 \times 227.1}{8} = -22.208$$

$$a_{22} = \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n} = 37421.12 - \frac{(544.8)^2}{8} = 320.24$$

$$a_{23} = \sum x_2 x_3 - \frac{(\sum x_2)(\sum x_3)}{n} = 15555.21 - \frac{544.8 \times 227.1}{8} = 89.70$$

$$a_{33} = \sum x_3^2 - \frac{\left(\sum x_3\right)^2}{n} = 6475.91 - \frac{(227.1)^2}{8} = 29.109$$

La matrice $\underset{\sim}{A}$ s'écrit donc

$$\underset{\sim}{A} = \begin{pmatrix} 194.6 & 12.18 & -22.208 \\ 12.18 & 320.24 & 89.7 \\ -22.208 & 89.7 & 29.109 \end{pmatrix}.$$

Pour obtenir la matrice $\underset{\sim}{S}$, il suffit de diviser la matrice $\underset{\sim}{A}$ par n–1 = 7. D'où on obtient

$$\underset{\sim}{S} = \begin{pmatrix} 27.799 & 1.74 & -3.1725 \\ 1.74 & 45.749 & 12.814 \\ -3,1725 & 12.814 & 4.1584 \end{pmatrix}.$$

Pour obtenir la matrice $\underset{\sim}{R}$, il faut calculer les corrélations entre les différentes

paires de variables en utilisant la formule (3.15).

$$r_{12} = \frac{1.74}{\sqrt{27.799 \times 45.749}} = 0.0488$$

$$r_{13} = \frac{-3.1725}{\sqrt{27.799 \times 4.1584}} = -0.2951$$

$$r_{23} = \frac{12.814}{\sqrt{45.749 \times 4.1584}} = 0.9291$$

Donc la matrice des corrélations s'écrit

$$\underset{\sim}{R} = \begin{pmatrix} 1.0 & 0.0488 & -0.2951 \\ 0.0488 & 1.0 & 0.9291 \\ -0.2951 & 0.9291 & 1.0 \end{pmatrix}$$

Iris de Fisher

Le calcul de la matrice des sommes de carrés et produits croisés, de la matrice de dispersion et de la matrice des corrélations pour les données des iris setosa de Fisher ($n = 50, p = 4$) conduit aux résultats suivants :

$$\underset{\sim}{\bar{x}}^T = (50.1, 34.3, 14.6, 2.46)$$

$$\underset{\sim}{A} = \begin{pmatrix} 608.825 & & & \\ 486.178 & 704.081 & & \\ 80.164 & 57.33 & 147.784 & \\ 50.617 & 45.57 & 29.743 & 54.439 \end{pmatrix}$$

$$\underset{\sim}{S} = \begin{pmatrix} 12.425 & & & \\ 9.922 & 14.369 & & \\ 1.636 & 1.170 & 3.016 & \\ 1.033 & 0.930 & 0.607 & 1.111 \end{pmatrix}$$

$$\underset{\sim}{R} = \begin{pmatrix} 1.00 & & & \\ 0.743 & 1.00 & & \\ 0.267 & 0.177 & 1.00 & \\ 0.278 & 0.233 & 0.332 & 1.00 \end{pmatrix}$$

Notons que les écarts-types des 4 variables s'obtiennent en prenant la racine carrée des éléments diagonaux de $\underset{\sim}{S}$. On a respectivement $s_1 = 3.52, s_2 = 3.79, s_3 = 1.74$ et $s_4 = 1.05$. Enfin, la matrice inverse de $\underset{\sim}{S}$ s'écrit

$$\underset{\sim}{S}^{-1} = \begin{pmatrix} 0.189 & & & \\ -0.124 & 0.156 & & \\ -0.045 & 0.011 & 0.388 & \\ -0.048 & -0.021 & -0.179 & 1.060 \end{pmatrix}$$

3.6 Distance de Mahalanobis

Considérons une observation multivariée $\underset{\sim}{x}^T = (x_1, \ldots, x_p)$. Pour rappel, la distance euclidienne (au carré) entre $\underset{\sim}{x}$ et le vecteur moyen $\underset{\sim}{\bar{x}}$ du nuage de points s'écrit

$$\begin{aligned} d^2(\underset{\sim}{x}) &= (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T(\underset{\sim}{x} - \underset{\sim}{\bar{x}}) \\ &= \sum_{j=1}^{p}(x_j - \bar{x}_j)^2 \end{aligned} \tag{3.19}$$

c'est-à-dire la somme des carrés des différences des composantes des deux vecteurs. Malheureusement, cette distance ne tient pas compte des associations entre les différentes variables.

Par définition, la *distance de Mahalanobis* entre $\underset{\sim}{x}$ et $\underset{\sim}{\bar{x}}$ est la forme quadratique

$$D^2(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T \underset{\sim}{S}^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}}) \tag{3.20}$$

où $\underset{\sim}{S}^{-1}$ est l'inverse de la matrice de dispersion. Dans le cas univarié ($p = 1$), les expressions (3.19) et (3.20) se réduisent respectivement aux formes suivantes

$$\begin{aligned} d^2(x) &= (x - \bar{x})^2 \\ D^2(x) &= \left(\frac{x - \bar{x}}{s}\right)^2 \end{aligned}$$

On constate que la distance de Mahalanobis tient compte de la dispersion des valeurs de l'échantillon (car s est l'écart-type), alors que ce n'est pas le cas pour la distance euclidienne.

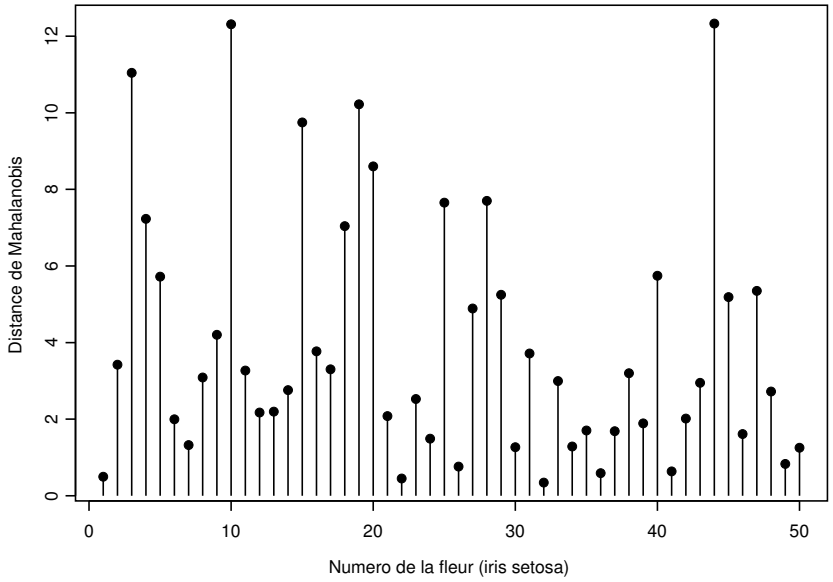
Dans le cas bivarié ($p = 2$), les points situés à une même distance euclidienne de la moyenne ($d^2 = c$) se trouvent sur un cercle, alors que pour la distance de Mahalanobis ($D^2 = c$), ils se trouvent sur une ellipse. De même à 3 dimensions ($p = 3$), au lieu d'une sphère, on a un ellipsoïde.

Si on calcule la distance de Mahalanobis pour toutes les observations multivariées $\underset{\sim}{x}_i$ ($i = 1, \ldots, n$) de l'échantillon, soit

$$D^2\left(\underset{\sim}{x}_i\right) = \left(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}\right)^T \underset{\sim}{S}^{-1}\left(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}\right),$$

on peut ordonner celles-ci par ordre croissant, de la plus petite à la plus grande. On voit ainsi les observations proches de la moyenne et celles qui s'en écartent fortement.

La figure ci-dessous reporte la distance de Mahalanobis en fonction du numéro de la fleur pour les iris setosa. On constate que certaines fleurs sont à une distance élevée du centre de la distribution. Ainsi la fleur 44 est la plus éloignée ($D^2 = 12.33$) alors que la fleur 32 est la plus proche ($D^2 = 0.34$).



Lorsque n est élevé, on sait que la distance de Mahalanobis se distribue comme une loi chi-carré à p degrés de liberté. On pourrait dès lors considérer qu'une observation est *extrême* ou *aberrante*, lorsque

$$D^2\left(\underset{\sim}{x}\right) \geq Q_{\mathcal{X}^2}(0.99; p) \tag{3.21}$$

où $Q_{\mathcal{X}^2}(0.99; p)$ est le quantile (percentile) à 99% du chi-carré à p degrés de liberté. Celui-ci n'est dépassé que par 1% des observations. On consulte à cet effet la table du chi-carré.

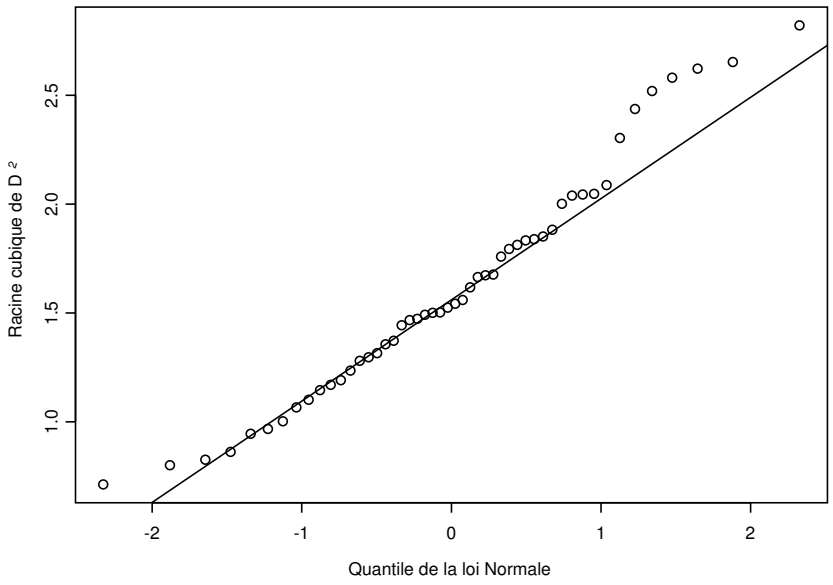
Lorsque n est petit, il est préférable d'utiliser le seuil critique suivant

$$D^2\left(\underset{\sim}{x}\right) \geq \frac{p(n^2-1)\; Q_F\;(0.99; p, n-p)}{n(n-p)} \tag{3.22}$$

où $Q_F(0.99; p, n-p)$ est le quantile à 99% du F de Snedecor à p et n–p degrés de liberté. A titre d'exemple, pour $n = 50$ et $p = 4$, dans le premier cas, le seuil critique vaut 13.28 et dans le second $(4 \times 2499 \times 3.76)/(50 \times 46) = 16.33$.

Les considérations précédentes doivent être prises avec précaution car s'il y a des observations extrêmes ou aberrantes dans l'échantillon multivarié, celles-ci affectent non seulement le vecteur moyen $\bar{x}$ et la matrice de dispersion $\underset{\sim}{S}$ mais aussi la distance de Mahalanobis.

Certains auteurs ont proposé d'étudier la distribution de $\sqrt[3]{D^2(x)}$, notamment en utilisant une échelle dite "anamorphose gaussienne" (ou de probabilité). Sur un tel graphique, les points devraient se distribuer sur une ligne droite. Des points isolés ou une courbure dans l'allure des points indiquent la présence de valeurs aberrantes ou la non normalité de la distribution des données. La figure ci-dessous illustre ces considérations pour les iris de Fisher.



3.7 Paradoxe de Rao

Dans le cas d'une variable qui suit une distribution normale, on définit généralement l'intervalle de référence comme celui qui contient 95% des individus de la population. Cet intervalle s'obtient en calculant $\bar{x} \pm 1.96s$. Une observation x est *anormale*, si elle tombe en-dehors de cet intervalle, donc si $|x - \bar{x}|/s \geq 1.96$.

Au niveau multivarié, quand on dispose de p variables $X_1, \ldots, X_p$, dont la distribution est normale, on peut déterminer les intervalles de référence pour chaque variable, soient

$$\bar{x}_i \pm 1.96 s_i \qquad (i = 1, \ldots, p)$$

où $\bar{x}_i$ et s_i sont respectivement la moyenne et l'écart-type de la variable X_i. Ces p intervalles de référence définissent une "boîte rectangulaire" dans

l'espace à p-dimensions $\mathbb{R}^p$. Un sujet $\underset{\sim}{x}$ est normal si $|x_i - \bar{x}_i|/s_i \leq 1.96$ $(i = 1, \ldots, p)$, c'est-à-dire s'il tombe à l'intérieur de la boîte rectangulaire.

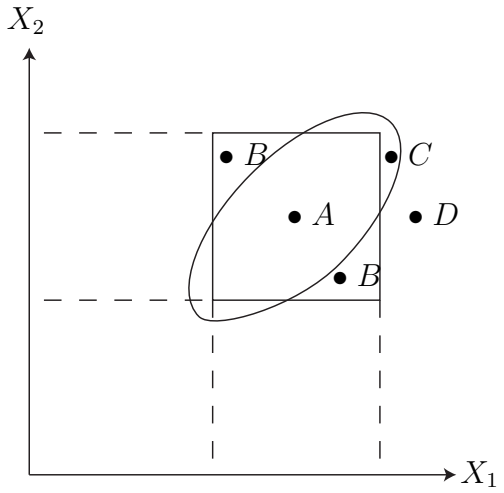
Cette approche ne tient pas compte des relations entre les différentes variables étudiées. Au contraire, il est préférable de définir une région de référence ellipsoïdale comme on l'a vu ci-dessus. Dans ces circonstances, un sujet est normal s'il tombe à l'intérieur de l'ellipsoïde qui contient 95% des sujets de la population, c'est-à-dire si

$$\left(\underset{\sim}{x} - \underset{\sim}{\bar{x}}\right)^T \underset{\sim}{S}^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\bar{x}}\right) \geq Q_{\mathcal{X}^2}(0.95; p)$$

et anormal s'il tombe en dehors de l'ellipse.

La confrontation du rectangle et de l'ellipse conduit au paradoxe de Rao, dans la mesure où un sujet peut tomber en dehors de l'ellipse (c'est-à-dire être "anormal" au niveau multivarié) mais à l'intérieur de la boîte rectangulaire (c'est-à-dire être "normal" pour chaque variable individuellement). Ceci s'explique par la prise en compte des associations entre les variables. Ainsi, si on considère le poids et la taille, un sujet peut avoir une taille normale et un poids normal mais la confrontation des deux mesures est "anormale" (excès de poids, maigreur). A l'inverse, un sujet peut avoir une taille et un poids excessifs mais les deux mesures biométriques restent en harmonie lorsqu'elles sont envisagées ensemble.

Comme on le voit sur la figure ci-après, plusieurs cas sont possibles : être normal au niveau univarié et multivarié (cas A), être normal au niveau univarié mais non au niveau multivarié (cas B), être anormal au niveau univarié mais normal au niveau multivarié (cas C), être anormal tant au niveau univarié que multivarié (cas D).



3.8 Remarque finale

Le vecteur moyen $\underset{\sim}{\bar{x}}$ et la matrice de variances-covariances $\underset{\sim}{S}$, de même que la matrice de corrélations $\underset{\sim}{R}$, peuvent être calculées quelles que soient les dimensions de la matrice d'observations $\underset{\sim}{X}_{n\times p}$. Toutefois, il faut faire remarquer que si $n \leq p$, c'est-à-dire si la matrice d'observations contient moins ou autant de sujets que de variables, les matrices $\underset{\sim}{S}$ et $\underset{\sim}{R}$ sont singulières. En d'autres termes, le déterminant est nul et les matrices ne peuvent être inversées. Il s'ensuit dans ce cas qu'on ne peut calculer la distance de Mahalanobis, puisque $\underset{\sim}{S}^{-1}$ n'existe pas.

En général, il est dès lors recommandé d'avoir plus d'observations que de variables. Certains auteurs préconisent d'avoir 10 fois plus de sujets que de variables, donc que $n/p \geq 10$ mais il s'agit d'une règle purement heuristique.

Indiquons enfin qu'en termes de population, la moyenne du vecteur $\underset{\sim}{X}$ s'écrit $\underset{\sim}{\mu}$ et la matrice de variances-covariances $\underset{\sim}{\sum}$. Comme en statistique univariée, les paramètres théoriques de population se désignent par des lettres grecques.

Chapitre 4

Analyse en composantes principales

4.1 Introduction

Ce chapitre est consacré à une méthode statistique multivariée importante, appelée *analyse en composantes principales (ACP)*. Il s'agit essentiellement d'une méthode descriptive relevant du domaine de la statistique exploratoire. On introduira d'abord la méthode à partir de la matrice de variances-covariances, ensuite on utilisera la matrice de corrélations. On terminera le chapitre en évoquant brièvement la méthode du *biplot*.

4.2 Objectifs

De tout temps, l'homme a cherché à obtenir une représentation bidimensionnelle d'un nuage de points de l'espace multivarié. L'esprit humain ne peut guère aller au-delà de 3 dimensions, celles dans lesquelles nous vivons. En relativité restreinte, on travaille dans un espace spatio-temporel à 4 dimensions, les 3 dimensions classiques et le temps.

L'analyse en composantes principales poursuit plusieurs objectifs.

1. Obtenir une représentation graphique (à 2 ou 3 dimensions) d'un échantillon multivarié, c'est-à-dire d'un nuage de points, avec une perte minimale d'information.
2. Mettre en évidence dans l'espace p-dimensionnel des observations atypiques ou extrêmes, voire aberrantes, ou encore des sous-groupes d'observations bien individualisés (appelés "clusters").

3. Trouver la véritable dimension du problème multivarié étudié, car en général les variables sont corrélées et contiennent dès lors des informations redondantes.
4. Réduire le nombre de variables étudiées en les remplaçant par un nombre plus limité de facteurs, qui sont des sommes pondérées des variables initiales. Ces facteurs peuvent alors être utilisés dans d'autres analyses statistiques.

La méthode du biplot permet en outre d'obtenir une représentation graphique des variables, ce qui donne une dimension supérieure à l'analyse exploratoire et permet de mieux comprendre les relations entre les variables.

4.3 Définition du problème

Soit un vecteur $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ de p variables et un échantillon simplement fortuit d'effectif n représenté par la matrice d'observations $\underset{\sim}{X}_{n\times p}$. Aucune restriction n'est faite sur les valeurs de n et de p. Il n'est pas exclu que $n < p$. Soient $\underset{\sim}{\bar{x}}$ et $\underset{\sim}{S}$, respectivement le vecteur moyen et la matrice de variances-covariances de l'échantillon, comme décrits au Chapitre 3. Ainsi, dans l'espace à p dimensions $\mathbb{R}^p$, l'échantillon forme un nuage de points dont $\underset{\sim}{\bar{x}}$ est le centre de gravité et dont $\underset{\sim}{S}$ traduit la forme et la dispersion.

Le problème statistique consiste à trouver un nouveau vecteur de variables $\underset{\sim}{Y}^T = (Y_1, \ldots, Y_p)$ à p composantes, qui présente les propriétés suivantes :

(i) les variables Y_i, appelées *composantes principales* (CP), sont des sommes pondérées des variables initiales

$$Y_i = \underset{\sim}{a}_i^T \underset{\sim}{X} = a_{1i}X_1 + \ldots + a_{pi}X_p \tag{4.1}$$

avec la condition que les coefficients (ou poids) soient normés à l'unité, c'est-à-dire $a_{1i}^2 + \ldots + a_{pi}^2 = 1$, ce qu'on écrit $||\underset{\sim}{a}_i|| = 1$ $(i = 1, \ldots, p)$

(ii) les variables Y_i ont des variabilités de plus en plus petites,

$$\text{var } (Y_1) \geq \text{ var } (Y_2) \geq \ldots \geq \text{ var } (Y_p) \geq 0 \tag{4.2}$$

(iii) les variables Y_i sont non corrélées,

$$\text{corr } (Y_i, Y_j) = 0 \quad \forall i \neq j \tag{4.3}$$

En général, on ne se sert que des deux premières composantes principales Y_1 et Y_2, les autres n'apportant souvent guère d'information complémentaire.

On passe ainsi de p dimensions à $q = 2$ dimensions, ce qui était l'objectif de la méthode.

Pour rappel, si on a un vecteur $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ à p variables et un vecteur de nombres $\underset{\sim}{a}^T = (a_1, \ldots, a_p)$, le produit scalaire $\underset{\sim}{a}^T \underset{\sim}{X} = a_1 X_1 + \ldots + a_p X_p$ donne une quantité univariée. Donc, le produit scalaire (ou somme pondérée) permet de passer d'un point de l'espace à p dimensions $\underset{\sim}{x}^T = (x_1, \ldots, x_p)$ à un point $y = \underset{\sim}{a}^T \underset{\sim}{x}$ sur la droite, c'est-à-dire un nombre.

4.4 Principe de la méthode

4.4.1 Première composante principale

Pour trouver Y_1, la première composante principale, il faut rechercher un vecteur de poids $\underset{\sim}{a_1}^T = (a_{11}, \ldots, a_{p1})$, normé à l'unité $||\underset{\sim}{a_1}|| = 1$, tel que la variabilité de la variable $Y_1 = \underset{\sim}{a_1}^T \underset{\sim}{X}$ soit la plus grande possible. En d'autres termes, si on calcule Y_1 pour chacune des n observations multivariées $\underset{\sim}{x_i}$ $(i = 1, \ldots, n)$, on obtient n nombres, notés

$$\left\{ y_{11} = \underset{\sim}{a_1}^T \underset{\sim}{x_1}, \ldots, y_{n1} = \underset{\sim}{a_1}^T \underset{\sim}{x_n} \right\} \tag{4.4}$$

dont la variance

$$V_1^2 = \sum_{k=1}^{n} (y_{k1} - \bar{y}_1)^2 / (n-1), \tag{4.5}$$

où $\bar{y}_1 = \sum_{k=1}^{n} y_{k1}/n$ est la moyenne des n nombres, doit être maximale.

Une autre façon de concevoir le problème, c'est de rechercher dans le nuage de points la direction où la variabilité est la plus grande. Imaginons une droite traversant le nuage de points et projetons tous les points du nuage sur cette droite. Les points sur la droite sont plus ou moins dispersés. Faisons alors bouger la droite dans le nuage de points dans toutes les directions jusqu'à ce que la dispersion des points sur la droite soit la plus grande possible. On trouve ainsi la première composante principale Y_1. C'est aussi la direction qui minimise la somme des carrés des distances entre chaque point et sa projection sur la droite (Pearson).

Pour trouver le vecteur $\underset{\sim}{a_1}$, il faut rechercher la plus grande valeur propre de la matrice de dispersion $\underset{\sim}{S}$, autrement dit il faut résoudre l'équation polynomiale.

$$|\underset{\sim}{S} - \lambda \underset{\sim}{I}| = 0 \tag{4.6}$$

Soient $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, les p solutions de ce polynôme d'ordre p classées par ordre décroissant. Dans ces conditions, $\underset{\sim}{a}_1$ est solution de l'équation

$$\underset{\sim}{\underset{\sim}{S}}\underset{\sim}{a}_1 = \lambda_1 \underset{\sim}{a}_1 \tag{4.7}$$

c'est-à-dire que $\underset{\sim}{a}_1$ est le vecteur propre associé à la plus grande valeur propre λ_1 de $\underset{\sim}{S}$ et normé à l'unité.

On a construit ainsi la première composante principale $Y_1 = \underset{\sim}{a}_1^T \underset{\sim}{X}$ et on trouve que $V_1^2 = \lambda_1$.

4.4.2 Deuxième composante principale

Pour trouver Y_2, la deuxième composante principale, il faut rechercher un vecteur de poids $\underset{\sim}{a}_2^T = (a_{12}, \ldots, a_{p2})$, normé à l'unité $\|\underset{\sim}{a}_2\| = 1$, tel que la variabilité de la variable $Y_2 = \underset{\sim}{a}_2^T \underset{\sim}{X}$ soit la plus grande possible après celle de Y_1 mais aussi que la corrélation entre Y_1 et Y_2 soit nulle.

Si on calcule Y_2 pour chacune des n observations multivariées $\underset{\sim}{x}_i$ ($i = 1, \ldots, n$), on obtient n nombres notés

$$\left\{ y_{12} = \underset{\sim}{a}_2^T \underset{\sim}{x}_1, \ldots, y_{n2} = \underset{\sim}{a}_2^T \underset{\sim}{x}_n \right\} \tag{4.8}$$

La variance de ces nombres, soit

$$V_2^2 = \sum_{k=1}^{n} (y_{k2} - \bar{y}_2)^2/(n-1) \tag{4.9}$$

où $\bar{y}_2 = \sum_{k=1}^{n} y_{k2}/n$ est la moyenne des n nombres, doit être maximale mais inférieure ou égale à V_1^2.

Par ailleurs, la corrélation entre Y_1 et Y_2, soit

$$r_{12} = \frac{\sum_{k=1}^{n} (y_{k1} - \bar{y}_1)(y_{k2} - \bar{y}_2)}{V_1 \cdot V_2}, \tag{4.10}$$

doit être nulle (sachant que $\underset{\sim}{a}_1$ a déjà été calculé !).

D'un point de vue géométrique, on recherche la deuxième direction la plus dispersée dans le nuage de points mais orthogonale (c'est-à-dire à angle droit) avec la première.

Pour trouver $\underset{\sim}{a}_2$, il faut résoudre l'équation

$$\underset{\sim}{\underset{\sim}{S}}\underset{\sim}{a}_2 = \lambda_2 \underset{\sim}{a}_2 \tag{4.11}$$

où λ_2 est la deuxième plus grande valeur propre de $\underset{\sim}{S}$. En conséquence, $\underset{\sim}{a}_2$ est le vecteur propre associé à λ_2 normé à l'unité et $V_2^2 = \lambda_2$. Par ailleurs, $r_{12} = 0$.

4.4.3 Représentation graphique

Les deux premières composantes principales, Y_1 et Y_2, définissent un système d'axes orthogonaux et donc un plan "principal" sur lequel on peut reporter les n observations multivariées $\underset{\sim}{x}_i$ $(i = 1, \ldots, n)$. Il suffit pour cela de reporter sur le plan les n points.

$$\begin{array}{c} (y_{11}, y_{12}) \\ (y_{21}, y_{22}) \\ \ldots \ \ldots \\ (y_{n1}, y_{n2}) \end{array} \tag{4.12}$$

définis en (4.4) et (4.8).

On obtient ainsi une représentation à 2 dimensions du nuage de points de l'espace $\mathbb{R}^P$. C'était un des objectifs de l'analyse.

4.4.4 Qualité de la représentation

Que vaut la "qualité" de la représentation à 2 dimensions du nuage de points car il y a forcément perte d'information (c'est le cas d'une photographie d'un objet tridimensionnel) ?

Le processus de construction des composantes principales peut se poursuivre au-delà des deux premières. Ainsi, pour trouver la troisième composante principale, $Y_3 = \underset{\sim}{a}_3^T \underset{\sim}{x}$, il faut résoudre l'équation $\underset{\sim}{S}\underset{\sim}{a}_3 = \lambda_3 \underset{\sim}{a}_3$, où λ_3 est la troisième plus grande valeur propre de $\underset{\sim}{S}$. On peut continuer ainsi de suite jusqu'à la p-ième et dernière composante principale $Y_p = \underset{\sim}{a}_p^T \underset{\sim}{X}$, où $\underset{\sim}{a}_p$ est solution de l'équation $\underset{\sim}{S}\underset{\sim}{a}_p = \lambda_p \underset{\sim}{a}_p$ et λ_p est la plus petite valeur propre de $\underset{\sim}{S}$. Si $\lambda_p = 0$, alors la composante principale Y_p est constante et elle n'a plus d'intérêt statistique puisque $V_p^2 = 0$.

Pour évaluer la qualité des deux premières composantes principales, Y_1 et Y_2, on procède comme suit.

La variabilité totale exprimée par les p composantes principales est égale à $\lambda_1 + \lambda_2 + \ldots + \lambda_p$. Or on sait que la somme des valeurs propres d'une matrice est égale à la trace de la matrice, c'est-à-dire à la somme de ses éléments diagonaux. Dès lors,

$$\lambda_1 + \ldots + \lambda_p = tr\ \underset{\sim}{S} = s_{11} + s_{22} + \ldots + s_{pp} \tag{4.13}$$

Attention, il ne faut pas confondre $s_{11} = s_1^2$ qui est la variance de X_1 et $V_1^2 = \lambda_1$ qui est la variance de Y_1; de même pour les autres éléments. La somme est la même mais non les termes individuels !

La contribution de la première composante principale à la réduction du problème multivarié peut être évaluée à partir du rapport

$$\frac{\lambda_1}{\lambda_1 + \ldots + \lambda_p} = \frac{\lambda_1}{\mathrm{tr}\underset{\sim}{S}} \tag{4.14}$$

L'expression (4.14) exprime la proportion de variabilité exprimée par Y_1 par rapport à la variabilité totale. Plus ce rapport est proche de 1, plus la composante Y_1 est importante et moins les autres composantes sont utiles. On peut faire de même pour chaque composante Y_i en calculant le rapport $\lambda_i/\mathrm{tr}\ \underset{\sim}{S}$.

La qualité des deux premières composantes principales réunies se mesure donc par le rapport

$$Q = \frac{\lambda_1 + \lambda_2}{\mathrm{tr}\ \underset{\sim}{S}} \tag{4.15}$$

et inversement la perte d'information peut s'écrire

$$\mathcal{P} = 1 - Q = \frac{\lambda_3 + \ldots + \lambda_p}{\mathrm{tr}\ \underset{\sim}{S}} \tag{4.16}$$

En pratique, on connaît donc la somme des valeurs propres avant de les calculer puisqu'elle est égale à $\mathrm{tr}\underset{\sim}{S}$. Les logiciels statistiques permettent de fixer le nombre q de composantes principales que l'on veut déterminer. S'il arrive qu'à partir d'un moment une valeur propre est nulle, toutes les suivantes le seront aussi. Ceci signifie qu'il y a trop de variables dans le problème multivarié et que certaines variables sont des combinaisons linéaires des autres. Par exemple, si X_1 est la longueur d'un rectangle, X_2, la largeur et $X_3 =$ le périmètre, une valeur propre sera nulle puisque $X_3 = 2(X_1 + X_2)$.

4.5 Interprétation des composantes principales

Les composantes principales Y_1 et Y_2 peuvent aussi s'écrire

$$\begin{aligned} Y_1 &= \underset{\sim}{a}_1^T(\underset{\sim}{X} - \underset{\sim}{\bar{x}}) \\ Y_2 &= \underset{\sim}{a}_2^T(\underset{\sim}{X} - \underset{\sim}{\bar{x}}) \end{aligned} \tag{4.17}$$

en retirant la moyenne de manière à obtenir une représentation graphique des n points dans le plan (Y_1, Y_2) centrée sur l'origine.

On est tenté de donner un sens aux composantes principales, de les interpréter. Il s'agit d'un exercice difficile, subjectif et périlleux. En toutes circonstances, il faut être prudent. Si les variables $X_1, \ldots, X_p$ sont des dimensions biométriques (par exemple, mensuration sur un individu ou un objet), alors Y_1 représente la taille des objets (size) alors que Y_2 représente la forme (shape) de ces objets.

Une aide à l'interprétation des composantes principales peut être obtenue en calculant la corrélation entre les composantes principales et les variables de départ. Pour rappel, les composantes principales sont non corrélées entre elles. On montre facilement que la corrélation entre la composante principale Y_j et la variable X_i est donnée par la formule

$$\text{corr}(X_i, Y_j) = \frac{a_{ij}\sqrt{\lambda_j}}{s_i}(i, j = 1 \ldots p) \tag{4.18}$$

où λ_j est la j-ème valeur propre de $\underset{\sim}{S}$, a_{ij} le coefficient de la variable X_i dans la composante principale Y_j et s_i l'écart-type de la variable X_i.

Ainsi, si on calcule la corrélation entre la première composante principale et chaque variable $X_1, \ldots, X_p$, on peut considérer que la composante principale représente surtout les variables avec lesquelles elle est fort corrélée (négativement ou positivement). Par contre, elle ne représente guère celles avec lesquelles la corrélation est faible.

L'analyse en composantes principales sur $\underset{\sim}{S}$ pose problème lorsqu'une ou plusieurs variables ont des variances fort différentes (facteur de 1 à 100). Les variables à forte variance l'emportent et les composantes principales surestiment le poids de ces variables, les autres variables à variance plus faible étant laissées pour compte. Si tel est le cas, il est préférable de faire une analyse en composantes principales sur la matrice des corrélations $\underset{\sim}{R}$.

4.6 ACP sur la matrice des corrélations

4.6.1 Vecteur centré réduit

L'analyse en composantes principales sur la matrice de corrélations $\underset{\sim}{R}$ repose sur le fait que la matrice $\underset{\sim}{R}$ est la matrice de variances-covariances du vecteur de variables centrées réduites $\underset{\sim}{Z}^T = (Z_1, \ldots, Z_p)$, où

$$Z_j = \frac{X_j - \bar{x}_j}{s_j} \quad (j = 1, \ldots, p) \tag{4.19}$$

ce que l'on écrit sous la forme matricielle

$$\underset{\sim}{Z} = \underset{\sim}{D}^{-1}(\underset{\sim}{X} - \underset{\sim}{\bar{x}}) \tag{4.20}$$

où $\underset{\sim}{D} = \text{diag}\ (s_1, \ldots, s_p)$ et $\underset{\sim}{\bar{x}}$ le vecteur moyen.

Donc, si on calcule la matrice de dispersion de la matrice d'observations centrées réduites

$$\underset{\sim}{Z}_{n\times p} = \begin{bmatrix} z_{11} & z_{12} & \ldots & z_{1p} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ z_{n1} & z_{n2} & \ldots & z_{np} \end{bmatrix} \tag{4.21}$$

où $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, on constate que $\underset{\sim}{S}_z = \underset{\sim}{R}$ et que par ailleurs $\underset{\sim}{\bar{z}} = \underset{\sim}{0}$.

4.6.2 Recherche des composantes principales

Pour trouver les composantes principales,

$$Y_j = \underset{\sim}{a}_j^T \underset{\sim}{Z} \quad (j = 1, \ldots, p) \tag{4.22}$$

on recherche les valeurs propres par ordre décroissant de la matrice $\underset{\sim}{R}$ et les vecteurs propres qui leur sont associés.

Soient à nouveau $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ les p valeurs propres de la matrice $\underset{\sim}{R}$, c'est-à-dire les solutions de l'équation polynomiale

$$|\underset{\sim}{R} - \lambda \underset{\sim}{I}| = 0 \tag{4.23}$$

La première composante principale Y_1 s'écrit dès lors

$$Y_1 = \underset{\sim}{a}_1^T \underset{\sim}{Z}$$

où $\underset{\sim}{a}_1$ est solution de l'équation matricielle $\underset{\sim}{R}\underset{\sim}{a}_1 = \lambda_1 \underset{\sim}{a}_1$.

La deuxième composante principale a pour expression $Y_2 = \underset{\sim}{a}_2^T \underset{\sim}{Z}$, où $\underset{\sim}{a}_2$ est solution de l'équation matricielle $\underset{\sim}{R}\underset{\sim}{a}_2 = \lambda_2 \underset{\sim}{a}_2$. On procède ainsi de suite jusqu'à l'arrêt du processus.

4.6.3 Qualité

La qualité des deux premières composantes principales s'écrit

$$Q = \frac{\lambda_1 + \lambda_2}{\text{tr}\underset{\sim}{R}} \tag{4.24}$$

puisque $\mathrm{tr}\underset{\sim}{R} = \lambda_1 + \ldots + \lambda_p$. Par ailleurs, puisque $\underset{\sim}{R}$ est la matrice des corrélations (avec des 1 sur la diagonale), $\mathrm{tr}\underset{\sim}{R} = p$, le nombre de variables du problème. Dès lors, l'expression (4.24) devient

$$Q = \frac{\lambda_1 + \lambda_2}{p} \tag{4.25}$$

et la perte d'information

$$\mathcal{P} = 1 - Q = \frac{\lambda_3 + \ldots + \lambda_p}{p} \tag{4.26}$$

4.6.4 Interprétation

Pour interpréter les composantes principales $Y_1, \ldots, Y_p$, on calcule à nouveau les corrélations avec les variables $X_1, \ldots, X_p$ en notant que $\mathrm{corr}(Y_j, X_i) = \mathrm{corr}(Y_j, Z_i)$ car Z_i n'est qu'une transformation linéaire de X_i. On a alors

$$\mathrm{corr}(Y_j, X_i) = a_{ij}\sqrt{\lambda_j} \quad (i, j = 1, \ldots, p) \tag{4.27}$$

Dans l'ACP sur la matrice $\underset{\sim}{R}$, toutes les variables sont mises sur le même pied d'égalité puisque leurs variances valent toutes 1 au travers de la transformation $\underset{\sim}{Z}$.

Il convient de faire remarquer qu'il n'y a pas de relation entre les composantes principales obtenues à partir de $\underset{\sim}{S}$ et celles obtenues à partir de $\underset{\sim}{R}$.

4.7 Exemple : les iris de Fisher

L'analyse en composantes principales est illustrée sur les données des iris setosa de Fisher (voir Annexe I). La matrice d'observations est constituée de $n = 50$ iris setosa pour lesquelles on a mesuré $p = 4$ variables : $X_1 =$ longueur des sépales, $X_2 =$ largeur des sépales, $X_3 =$ longueur des pétales et $X_4 =$ largeur des pétales. On dispose donc d'une matrice d'observations de dimension $n \times p = 50 \times 4$.

4.7.1 ACP sur la matrice de variances-covariances

Présentons d'abord les résultats de l'ACP sur la matrice de variances-covariances $\underset{\sim}{S}$. Celle-ci s'écrit comme on l'a vu au paragraphe 3.5.

$$\underset{\sim}{S} = \begin{pmatrix} 12.425 & & & \\ 9.922 & 14.369 & & \\ 1.636 & 1.170 & 3.016 & \\ 1.033 & 0.930 & 0.607 & 1.111 \end{pmatrix}.$$

On note que la trace de la matrice $\underset{\sim}{S}$ vaut $tr\underset{\sim}{S} = 30.92$. Les valeurs et vecteurs propres de $\underset{\sim}{S}$ sont donnés ci-dessous. On constate que la première valeur propre $\lambda_1 = 23.65$ compte pour plus de 75% de la variance expliquée. Les deux premières valeurs propres donnent une représentation bidimensionnelle des observations de l'espace à 4 dimensions avec une qualité de 88.4%.

Valeur propre	Qualité (%)	Qualité cumulée
$\lambda_1 = 23.65$	76.47	76.47
$\lambda_2 = 3.69$	11.94	88.41
$\lambda_3 = 2.68$	8.67	97.08
$\lambda_4 = 0.90$	2.92	100.0

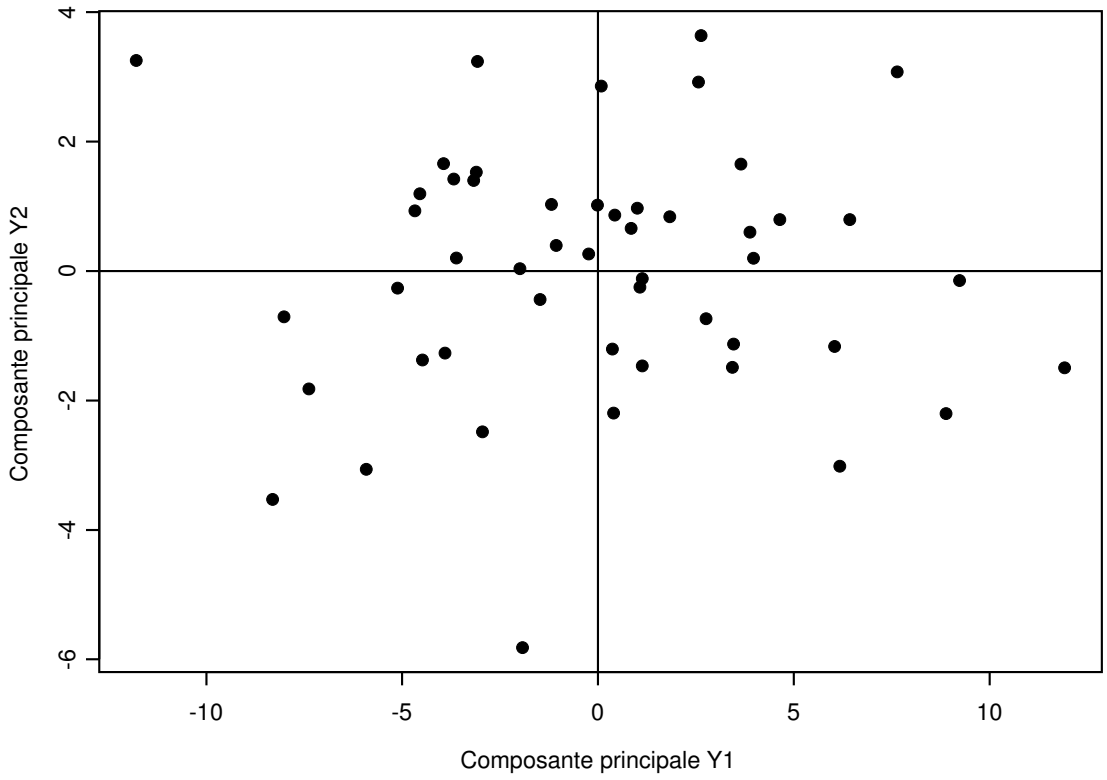
	Vecteur propre			
Variable	$\underset{\sim}{a}_1$	$\underset{\sim}{a}_2$	$\underset{\sim}{a}_3$	$\underset{\sim}{a}_4$
X_1	0.669	-0.598	0.440	0.036
X_2	0.734	0.621	-0.275	0.020
X_3	0.097	-0.490	-0.832	0.234
X_4	0.0636	-0.131	-0.195	-0.970

Ainsi la première composante principale s'écrit $Y_1 = 0.669(X_1 - 50.1) + 0.734(X_2 - 34.3) + 0.097(X_3 - 14.6) + 0.0636(X_4 - 2.46)$ et la seconde $Y_2 = -0.598(X_1 - 50.1) + 0.621(X_2 - 34.3) - 0.490(X_3 - 14.6) - 0.131(X_4 - 2.46)$.

La matrice des corrélations entre les variables et les composantes principales est reprise ci-après. On note que la première composante principale est corrélée positivement avec les 4 variables mais représente principalement les sépales. Par contre, la deuxième composante principale oppose la largeur des pétales aux autres variables.

Corrélations entre les variables et les composantes principales				
	Y_1	Y_2	Y_3	Y_4
X_1	0.923	0.326	-0.204	-0.010
X_2	0.942	-0.315	0.119	-0.005
X_3	0.270	0.542	0.785	-0.131
X_4	0.293	0.239	0.303	0.875

La figure ci-dessous donne la représentation des 50 iris setosa dans l'espace à deux dimensions des composantes principales Y_1 et Y_2. Rappelons qu'on a retiré la moyenne de chacune des quatre variables de départ afin que le nuage de points soit centré sur l'origine.



4.7.2 ACP sur la matrice de corrélations

Les résultats de l'ACP sur la matrice de corrélation $\underset{\sim}{R}$ sont repris ci-après. Cette matrice s'écrit (voir paragraphe 3.5)

$$\underset{\sim}{R} = \begin{pmatrix} 1.00 & & & \\ 0.743 & 1.00 & & \\ 0.267 & 0.177 & 1.00 & \\ 0.278 & 0.233 & 0.332 & 1.00 \end{pmatrix}$$

La première composante principale représente 51% de la variabilité totale puisque $tr\underset{\sim}{R} = 4$, alors que les deux premières composantes représentent plus de 75%. Cette qualité autorise une représentation satisfaisante des iris setosa dans le plan.

Valeur propre	Qualité (%)	Qualité cumulée
$\lambda_1 = 2.06$	51.46	51.46
$\lambda_2 = 1.02$	25.55	77.02
$\lambda_3 = 0.67$	16.70	93.71
$\lambda_4 = 0.25$	6.29	100.00

	Vecteur propre			
Variable	$\underset{\sim}{a}_1$	$\underset{\sim}{a}_2$	$\underset{\sim}{a}_3$	$\underset{\sim}{a}_4$
X_1	0.604	-0.335	0.067	0.720
X_2	0.576	-0.441	0.001	-0.689
X_3	0.375	0.627	0.677	-0.087
X_4	0.403	0.548	-0.733	-0.015

La matrice des corrélations entre les variables et les composantes principales est donnée ci-après.

Corrélations entre les variables et les composantes principales				
	Y_1	Y_2	Y_3	Y_4
X_1	0.867	-0.339	0.055	0.361
X_2	0.826	-0.446	0.001	-0.345
X_3	0.539	0.634	0.553	-0.044
X_4	0.578	0.554	-0.599	-0.007

La première composante principale s'écrit

$$Y_1 = 0.604\ Z_1 + 0.576\ Z_2 + 0.375\ Z_3 + 0.403\ Z_4,$$

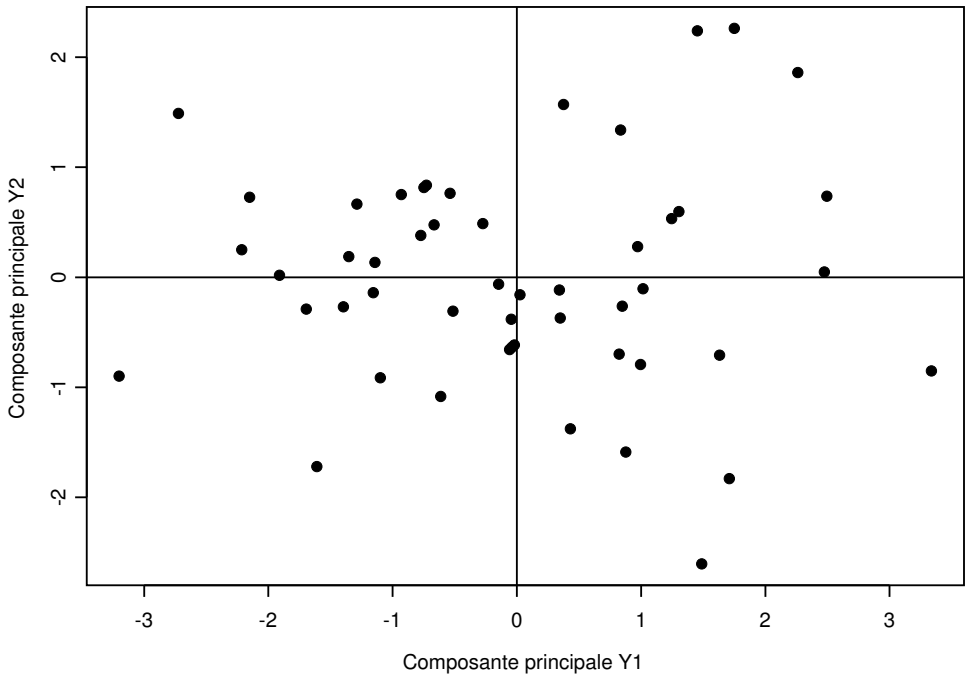
où $Z_1 = (X_1 - 50.1)/3.52, Z_2 = (X_2 - 34.3)/3.79, Z_3 = (X_3 - 14.6)/1.74$ et $Z_4 = (X_4 - 2.46)/1.05$.

Elle représente une somme pondérée des quatre variables dont les coefficients sont tous positifs. Elle correspond à un paramètre de "taille" (size). La deuxième composante principale

$$Y_2 = -0.335\ Z_1 - 0.441\ Z_2 + 0.627\ Z_3 + 0.548\ Z_4$$

est un "contraste" entre les pétales et les sépales. Elle oppose sépales et pétales et correspond à un paramètre de "forme" (shape).

La figure ci-après représente les 50 iris setosa dans le plan des deux premières composantes principales. L'abscisse représente la dimension des fleurs. En passant de gauche à droite sur le graphique, les iris ont des tailles de plus en plus grandes. Par contre, sur l'ordonnée, en passant du bas vers le haut, les iris ont des sépales de plus en plus petites mais des pétales de plus en plus grandes.



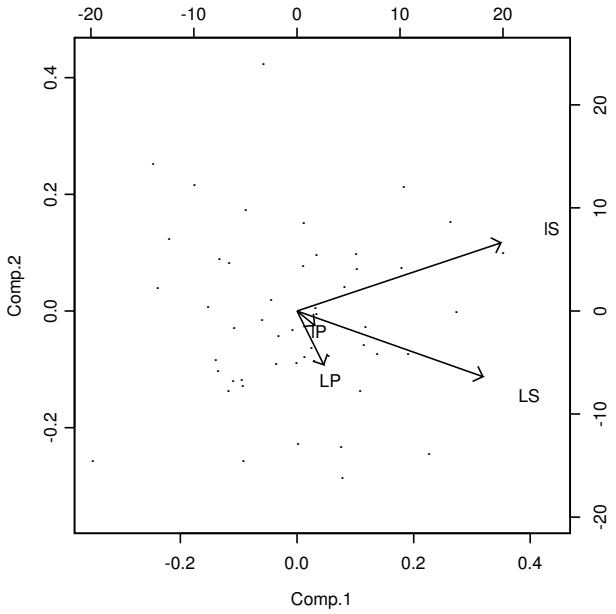
4.8 Biplot

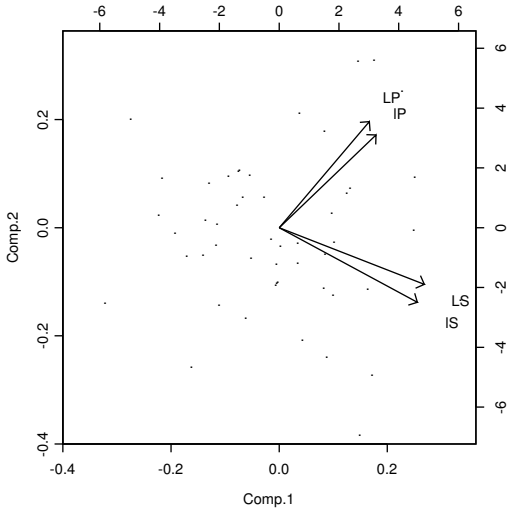
La méthode du biplot permet de représenter sur un même plan à la fois les n observations et les p variables. On a donc un double système d'axes, d'où le nom de *biplot*. Le développement théorique dépasse le cadre de ce cours. Chaque variable est représentée par un vecteur partant de l'origine du système d'axes et aboutissant au point où se situe la variable. Le principe d'interprétation est le suivant :

- des variables qui pointent dans la même direction sont corrélées positivement ;
- des variables qui pointent dans des directions diamétralement opposées sont corrélées négativement ;
- des variables qui pointent dans des directions orthogonales (à angle droit) ne sont pas corrélées.

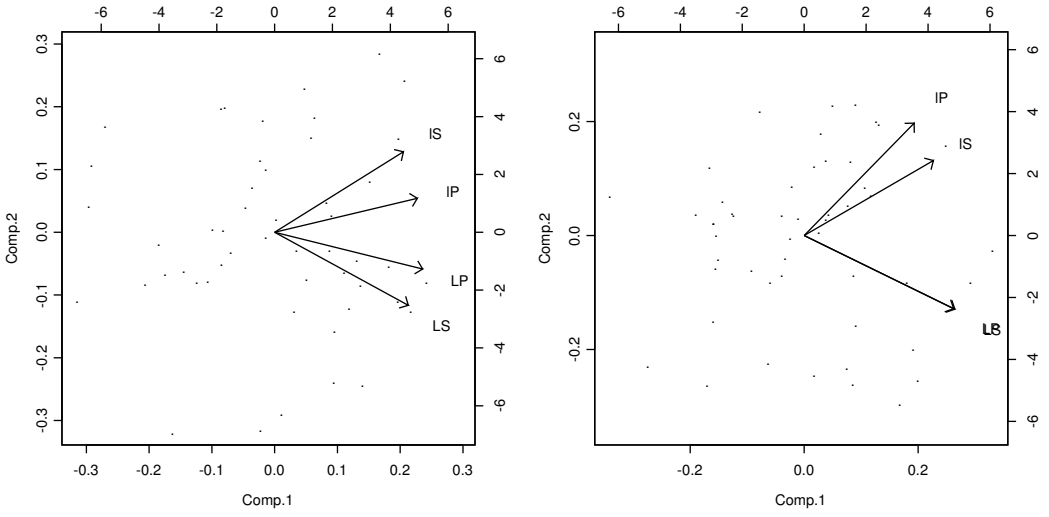
Par ailleurs, le sens du vecteur indique que l'on passe de valeurs basses à des valeurs élevées de la variable. Quant à la longueur du vecteur, elle représente l'importance de la variable. La méthode du biplot peut se faire sur la matrice $\underset{\sim}{S}$ ou sur la matrice $\underset{\sim}{R}$. Cette dernière est souvent préférée pour les mêmes raisons que l'ACP.

Les figures ci-dessous illustrent l'application du biplot sur les iris setosa de Fisher en se basant sur les matrices $\underset{\sim}{S}$ et $\underset{\sim}{R}$, respectivement.





On observe que les variables liées aux sépales (largeur et longueur) sont fortement corrélées entre elles car elles pointent dans la même direction. Il en est de même pour la longueur et la largeur des sépales. Par contre, sépales et pétales sont orthogonales et donc sans liaison entre elles. De façon étonnante, les biplots basés sur la matrice R pour les iris versicolor et virginica (voir les deux figures ci-dessus) montrent que la longueur des sépales est fortement corrélée à la longueur des pétales et qu'il en est de même pour les largeurs entre elles. Par contre, les longueurs ne sont pas corrélées avec les largeurs.



Chapitre 5

Régression et corrélation multiple

5.1 Introduction

La méthode de régression multiple est l'une des plus utilisées en statistique. Elle n'est en général pas considérée comme faisant partie de la statistique multivariée mais davantage comme une technique visant à étudier la relation (l'association) entre la moyenne d'une variable aléatoire et différents facteurs expérimentaux dont les valeurs sont fixées par l'investigateur. Par exemple, dans une expérience de laboratoire, on peut fixer le pH, la température et la concentration d'un milieu à différentes valeurs et mesurer à chaque fois la durée d'une réaction chimique. Les conditions expérimentales sont fixées mais la durée de la réaction est observée, dans la mesure où sa valeur n'est pas connue à l'avance. Dans ce chapitre, nous envisageons la régression multiple sous l'aspect multivarié car elle fait appel au calcul matriciel. On étudiera ses propriétés et les tests d'hypothèses qui sont associés. Lorsque toutes les variables sont observées simultanément, on parle davantage d'un problème de corrélation multiple. On terminera le chapitre par quelques mots sur les méthodes de sélection de variables.

5.2 Définition du problème

Soit un vecteur de $p+1$ variables $(Y, \underset{\sim}{X}^T) = (Y, X_1, \ldots, X_p)$ où Y est appelée la variable "dépendante" et $\underset{\sim}{X}$ le vecteur des variables "indépendantes". Cette dernière dénomination est malencontreuse car les variables $X_1, \ldots, X_p$ ne sont pas indépendantes entre elles, au contraire elles sont souvent corrélées.

Le qualificatif "explicatives" convient mieux aux variables du vecteur $\underset{\sim}{X}$. On les appelle aussi facteurs ou critères !

On suppose que la variable Y est quantitative (continue) et que sa distribution suit une loi Normale (loi de Gauss). Aucune restriction n'est faite quant au nombre et à la nature des variables explicatives $X_1, \ldots, X_p$.

Deux situations distinctes peuvent se présenter :

1. Les valeurs du vecteur $\underset{\sim}{X}$ sont fixées et la variable Y est observée. On parle alors de *régression multiple*.
2. La variable Y et le vecteur $\underset{\sim}{X}$ sont observés simultanément. On parle alors de *corrélation multiple* mais aussi de régression multiple.

Dans les deux cas, les données se présentent sous la forme d'une matrice d'observations $\underset{\sim}{X}_{n\times(p+1)}$.

$$\underset{\sim}{X}_{n\times(p+1)} = \begin{bmatrix} y_1 & x_{11} & \ldots & x_{1p} \\ y_2 & x_{21} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ y_n & x_{n1} & \ldots & x_{np} \end{bmatrix} \tag{5.1}$$

5.3 Régression multiple

5.3.1 Définition du modèle

Considérons d'abord la première situation. On recherche une relation entre la variable Y et les facteurs explicatifs $X_1, \ldots, X_p$. Comme souvent, en statistique, on fait l'hypothèse d'un modèle linéaire, soit

$$E(Y|\underset{\sim}{X} = \underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

ou

$$E(Y|\underset{\sim}{x}) = \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} \tag{5.2}$$

où $E(Y|\underset{\sim}{X} = \underset{\sim}{x})$ est la moyenne de la variable Y pour une valeur $\underset{\sim}{X} = \underset{\sim}{x}$ fixée, $\underset{\sim}{\beta}^T = (\beta_1, \ldots, \beta_p)$ est le vecteur des coefficients de régression et β_0 l'ordonnée à l'origine (appelée aussi "intercept"). On a pris la moyenne de Y car il est impensable que toutes les valeurs de Y pour un $\underset{\sim}{X}$ donné soient sur le plan.

L'équation (5.2) est celle d'un hyperplan de l'espace à $p+1$ dimensions $\mathbb{R}^{p+1}$. Lorsque $p = 1$, on a une droite $E(Y|x) = \beta_0 + \beta_1 x$ et lorsque $p = 2$ on a l'équation d'un plan $E(Y|\underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

L'équation (5.2) est la "Régression multiple de Y sur $\underset{\sim}{X}$".

Si on connaissait les coefficients $\beta_0, \beta_1, \ldots, \beta_p$, alors, pour tout $\underset{\sim}{X} = \underset{\sim}{x}$ donné, on pourrait calculer la valeur moyenne de Y et ainsi obtenir une prédiction de Y pour un $\underset{\sim}{x}$ donné.

En termes de valeurs de Y, l'équation (5.2) peut s'écrire

$$Y|\underset{\sim}{x} = \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} + \epsilon \tag{5.3}$$

où ϵ est une erreur "aléatoire" de moyenne 0 et de variance égale à σ^2 pour tout $\underset{\sim}{x}$. La partie "$\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}$" est la composante "fixe" (calculable) du modèle et ϵ la partie aléatoire (imprévisible).

En se référant à la matrice d'observations (5.1), l'équation (5.3) peut aussi s'écrire :

$$\begin{array}{rcll} y_i & = & \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x_i} + \epsilon_i & (i = 1, \ldots, n) \\ & = & \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i & \end{array} \tag{5.4}$$

où $\underset{\sim}{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$ est le vecteur des erreurs (ou résidus), c'est-à-dire des écarts entre les observations y_i et le modèle $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x_i}$ puisque

$$\epsilon_i = y_i - (\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x_i}) \qquad (i = 1, \ldots, n) \tag{5.5}$$

Le modèle linéaire (5.2) contient $p + 1$ inconnues qui sont les paramètres du modèle, à savoir $\beta_0, \beta_1, \ldots, \beta_p$. Il convient de les estimer à partir de la matrice d'observations (5.1).

5.3.2 Estimation des paramètres

Pour estimer les paramètres du modèle (5.2), on utilise le *principe des moindres carrés*. En clair, on recherche les valeurs de β_0 et $\underset{\sim}{\beta}$ qui minimisent la somme des carrés des résidus, soit $\underset{\sim}{\epsilon}^T \underset{\sim}{\epsilon} = \epsilon_1^2 + \ldots + \epsilon_n^2$.

En vertu de l'équation (5.5), on recherche sur l'ensemble des valeurs de $\beta_0, \underset{\sim}{\beta}$ le minimum de la fonction

$$\sum_{i=1}^{n} \left[y_i - (\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x_i}) \right]^2 \tag{5.6}$$

Il suffit de dériver cette fonction par rapport à tous les paramètres et d'égaler la dérivée à 0. Toutefois, si on dérive par rapport à β_0, on obtient

$$\sum_{i=1}^{n} \left[y_i - (\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x_i}) \right] = 0$$

ou encore

$$\sum_{i=1}^{n} y_i - n\beta_0 - \underset{\sim}{\beta}^T \sum_{i=1}^{n} \underset{\sim}{x}_i = 0$$

En divisant chaque terme par n, on constate que la solution β_0 doit satisfaire à l'équation

$$\beta_0 = \bar{y} - \underset{\sim}{\beta}^T \underset{\sim}{\bar{x}} \tag{5.7}$$

En remplaçant dans l'expression (5.6), β_0 par sa valeur (5.7), on obtient la fonction

$$\sum_{i=1}^{n} \left[(y_i - \bar{y}) - \underset{\sim}{\beta}^T (\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}) \right]^2 \tag{5.8}$$

La dérivée de la fonction (5.8) par rapport au vecteur $\underset{\sim}{\beta}$ conduit à l'équation matricielle

$$\underset{\sim}{A} \cdot \underset{\sim}{\beta} = \underset{\sim}{c} \tag{5.9}$$

où $\underset{\sim}{A}$ est la matrice des sommes de carrés et produits croisés du vecteur $\underset{\sim}{X}$ (voir Equation (3.12)) et $\underset{\sim}{c}$ est le vecteur des sommes de produits croisés de Y avec chaque variable du vecteur $\underset{\sim}{X}$. En d'autres termes, $\underset{\sim}{A} = (n-1)\,\underset{\sim}{S}$, où $\underset{\sim}{S}$ est la matrice de dispersion du vecteur $\underset{\sim}{X}$ et $\underset{\sim}{c}^T = (c_1, \ldots, c_p)$ où $c_j = \sum_{i=1}^{n} (y_i - \bar{y})(x_{ij} - \bar{x}_j)$.

En pratique, il faut donc calculer $\sum x_i, \sum x_i^2, \sum x_i x_j$, comme pour la matrice $\underset{\sim}{S}$ mais aussi $\sum y, \sum y^2$ et $\sum y x_i$. Dans ces conditions, les éléments de la matrice $\underset{\sim}{A}$ et du vecteur $\underset{\sim}{c}$ valent respectivement

$$a_{ij} = \sum x_i x_j - \frac{(\sum x_i)\,(\sum x_j)}{n}$$

et (5.10)

$$c_j = \sum y x_j - \frac{(\sum y)\,(\sum x_j)}{n}$$

$(i, j = 1, \ldots, p)$.

Dès lors, la solution $\underset{\sim}{\hat{\beta}}$ ou b de l'équation (5.9) s'obtient en inversant la matrice $\underset{\sim}{A}$ et en calculant l'expression

$$\underset{\sim}{b} = \underset{\sim}{A}^{-1} \underset{\sim}{c} \tag{5.11}$$

En utilisant la relation (5.7), le terme indépendant vaut

$$b_0 = \bar{y} - \underset{\sim}{b}^T \underset{\sim}{\bar{x}} \tag{5.12}$$

Dans ces conditions, la régression multiple estimée s'écrit

$$\hat{Y} = b_0 + \underset{\sim}{b}^T \underset{\sim}{x} \tag{5.13}$$

5.3.3 Analyse de la variance

En régression multiple, on peut décomposer la variabilité totale des observations de la variable Y en deux parties : l'une liée au modèle (hyperplan) et l'autre à la variabilité des points autour du plan (résidus).

La somme des carrés totale s'écrit respectivement, en utilisant (5.13) et en posant $\hat{Y}_i = b_0 + \underset{\sim}{b}^T \underset{\sim i}{x}$,

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{Y}_i + \hat{Y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2 \end{aligned} \tag{5.14}$$

car le double produit s'annule par sommation.

Le terme de gauche est la *somme des carrés totale*

$$\mathrm{SCT} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \tag{5.15}$$

Le second terme de droite est la *somme des carrés due à la régression*

$$\mathrm{SCR} = \underset{\sim}{b}^T \underset{\sim}{c} = b_1 c_1 + \ldots + b_p c_p \tag{5.16}$$

Enfin, le premier terme de droite est la *somme des carrés résiduelle* ou somme de carrés de l'erreur qu'on écrit SCE. On a

$$\mathrm{SCE} = \mathrm{SCT} - \mathrm{SCR} \tag{5.17}$$

Les degrés de liberté associés aux trois sommes de carrés valent respectivement $n-1$ pour SCT, p pour SCR et $n-p-1$ pour SCE. En divisant les sommes de carrés par leurs degrés de liberté, on obtient les carrés moyens. En particulier,

$$\mathrm{CMR} = \mathrm{SCR/p} \tag{5.18}$$

est le *carré moyen dû à la régression*, c'est-à-dire la part de variabilité expliquée par le modèle (5.2), et

$$s^2 = \mathrm{SCE}/(n-p-1) \tag{5.19}$$

est la *variabilité résiduelle* non expliquée par le modèle. Il s'agit de l'estimation de σ^2 défini dans le modèle linéaire et qui mesure la variabilité

des observations $y_1, \ldots, y_n$ autour de l'hyperplan. Si toutes les observations étaient situées sur l'hyperplan, alors s^2 serait égal à 0.

Lorsque $\underset{\sim}{\beta} = \underset{\sim}{0}$, c'est-à-dire lorsque la moyenne de Y ne dépend pas du vecteur $\underset{\sim}{X}$, le rapport

$$F = \frac{\text{CMR}}{s^2} \tag{5.20}$$

est distribué comme un F de Snedecor à p et $n-p-1$ degrés de liberté.

En conséquence, si on teste l'hypothèse nulle $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$ versus l'hypothèse alternative $H_1 : \underset{\sim}{\beta} \neq \underset{\sim}{0}$, on rejette H_0 au niveau d'incertitude α si

$$F \geq Q_F(1-\alpha; p, n-p-1) \tag{5.21}$$

où $Q_F(1-\alpha; p, n-p-1)$ est la quantile $1-\alpha$ de la distribution du F de Snedecor à p et $n-p-1$ degrés de liberté.

La table d'analyse de la variance s'écrit donc

Variabilité	Somme de carrés	Degrés de liberté	Carré moyen	F
Régression	SCR	p	CMR	$F = \dfrac{\text{CMR}}{s^2}$
Résiduelle	SCE	$n-p-1$	s^2	
Totale	SCT	$n-1$		

En conclusion, l'analyse de la variance permet de voir si la régression multiple a un sens, c'est-à-dire de montrer que $\underset{\sim}{\beta}$ n'est pas nul. Attention $\underset{\sim}{\beta} \neq \underset{\sim}{0}$ ne signifie pas que $\beta_i \neq 0$ pour $\forall i$.

5.3.4 Utilité des variables explicatives

Si l'hypothèse nulle $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$ est rejetée, on peut s'interroger sur l'utilité de chacune des p variables explicatives du modèle.

Pour tester l'hypothèse $H_0 : \beta_i = 0$ vers $H_1 : \beta_i \neq 0$ $(i = 1, \ldots, p)$, il suffit de calculer le critère

$$t = \frac{b_i}{\text{SE}(b_i)} \tag{5.22}$$

distribué sous H_0 comme un t de Student à $n-p-1$ degrés de liberté. Dans cette expression, l'erreur type de b_i est donnée par l'expression

$$\text{SE}(b_i) = s\sqrt{a^{ii}} \tag{5.23}$$

où $s = \sqrt{s^2}$ et $a^{ii} = (\underset{\sim}{A}^{-1})_{ii}$, l'élément diagonal de la matrice inverse utilisée dans (5.11). En effet, on montre aisément que var $\underset{\sim}{b} = s^2 \underset{\sim}{A}^{-1}$.

On rejette $H_0 : \beta_i = 0$ si

$$|t| \geq Q_t(1 - \frac{\alpha}{2}; n - p - 1) \tag{5.24}$$

le quantile $1 - \frac{\alpha}{2}$ du t de Student à $n - p - 1$ degrés de liberté, sinon on ne rejette pas H_0.

En cas de non rejet de l'hypothèse nulle, on peut supposer $\beta_i = 0$, ce qui signifie que la variable X_i n'intervient pas dans le modèle. Celui-ci peut alors être simplifié.

5.3.5 Qualité de la régression multiple

La qualité du modèle de régression multiple (5.2) peut être appréciée à l'aide du critère (voir table ANOVA)

$$R^2 = \frac{\text{SCR}}{\text{SCT}} \tag{5.25}$$

appelé *coefficient de détermination multiple*. On a toujours $0 \leq R^2 \leq 1$. Plus R^2 est proche de 1, meilleur est la qualité de la régression car SCR est voisin de SCT, ce qui signifie que SCE ≈ 0.

La quantité R^2 représente la proportion (ou le pourcentage) de la variabilité des observations de Y expliquée par les variables explicatives $X_1, \ldots, X_p$.

5.3.6 Précision de la prédiction

Ayant estimé les paramètres du modèle, on peut estimer Y pour une valeur donnée $\underset{\sim}{X} = \underset{\sim}{x}$ en utilisant l'équation (5.13), soit $\hat{Y} = b_0 + \underset{\sim}{b}^T \underset{\sim}{x} = \bar{y} + \underset{\sim}{b}^T (\underset{\sim}{x} - \underset{\sim}{\bar{x}})$. La précision de cette estimation, c'est-à-dire son erreur type SE $(\hat{Y})$, est la racine carrée de l'expression

$$\text{var } \hat{Y} = s^2 \left\{ \frac{1}{n} + (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T \underset{\sim}{A}^{-1} (\underset{\sim}{x} - \underset{\sim}{\bar{x}}) \right\} \tag{5.26}$$

On peut donc calculer l'intervalle de confiance à 95% pour $E(Y|\underset{\sim}{x})$, soit

$$\hat{Y} \pm 1.96 \text{ SE}(\hat{Y}) \tag{5.27}$$

On constate aisément que cet intervalle de confiance est minimum en $\underset{\sim}{X} = \underset{\sim}{\bar{x}}$ puisque dans ce cas le second terme de (5.26) s'annule et on a SE $(\hat{Y}) = s/\sqrt{n}$.

5.4 Corrélation multiple

5.4.1 Définition

Supposons que Y et $\underset{\sim}{X}$ soient observés simultanément, signifiant qu'à la fois la variable Y et le vecteur $\underset{\sim}{X}$ sont aléatoires. Dans ces conditions, plus aucune variable n'est réellement privilégiée.

Rien n'empêche de calculer comme précédemment la régression de Y sur $\underset{\sim}{X}$, soit

$$E(Y|\underset{\sim}{x}) = \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}.$$

Cette expression définit une nouvelle variable, notée $\hat{Y} = E(Y|\underset{\sim}{x})$, qui est aussi la prédiction linéaire de Y à partir de $\underset{\sim}{X} = \underset{\sim}{x}$.

La corrélation classique entre la variable Y et la variable $\hat{Y}$ (elle-même dérivée du vecteur $\underset{\sim}{X}$) est appelée *corrélation multiple* entre Y et $\underset{\sim}{X}$.

Par définition, le coefficient de corrélation multiple s'écrit

$$R = \text{corr}\,(Y, \hat{Y}) \tag{5.28}$$

En pratique, à partir de la matrice d'observations (5.1), on a estimé la régression multiple de Y sur $\underset{\sim}{X}$, soit $\hat{Y} = b_0 + \underset{\sim}{b}^T \underset{\sim}{x}$. On peut donc calculer $\hat{Y}$ pour chaque observation et définir ainsi la matrice des valeurs observées et prédites de Y :

$$\begin{pmatrix} y_1 & \hat{Y}_1 \\ y_2 & \hat{Y}_2 \\ \cdots & \cdots \\ y_n & \hat{Y}_n \end{pmatrix} \tag{5.29}$$

La corrélation entre ces deux séries d'observations s'écrit, puisque $\bar{\hat{Y}} = \bar{y}$,

$$\hat{R} = \text{corr}\,(y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{Y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2}} \tag{5.30}$$

On montre facilement que le numérateur de (5.30) et que le 2e terme du dénominateur sous la racine carrée valent chacun $\underset{\sim}{b}^T \underset{\sim}{c}$. On peut donc écrire,

$$\hat{R} = \frac{\underset{\sim}{b}^T \underset{\sim}{c}}{\sqrt{\sum(y - \bar{y})^2 \cdot \underset{\sim}{b}^T \underset{\sim}{c}}} = \sqrt{\frac{\underset{\sim}{b}^T \underset{\sim}{c}}{\sum(y - \bar{y})^2}}$$

En se rappelant que SCT $= \sum(y - \bar{y})^2$ et SCR $= \underset{\sim}{b}^T \underset{\sim}{c}$, on a finalement

$$\hat{R} = \sqrt{\frac{\text{SCR}}{\text{SCT}}} \tag{5.31}$$

Par définition, $0 \leq \hat{R} \leq 1$. Plus le coefficient de corrélation multiple est proche de 1, meilleure est la corrélation entre Y et $\underset{\sim}{X}$.

5.4.2 Test d'hypothèse

La question qui se pose est de savoir si le coefficient de corrélation multiple entre Y et $\underset{\sim}{X}$ est nul, auquel cas Y et $\underset{\sim}{X}$ ne sont pas corrélés linéairement.

Pour tester l'hypothèse nulle $H_0 : R = 0$ versus $H_1 : R \neq 0$, il faut supposer que la distribution conjointe de $(Y, \underset{\sim}{X})$ est multinormale, donc aussi que les composantes du vecteur $\underset{\sim}{X}$ sont quantitatives et normales.

Sous H_0, le critère

$$F = \frac{n - p - 1}{p} \frac{\hat{R}^2}{1 - \hat{R}^2} \tag{5.32}$$

est distribué comme un F de Snedecor à p et $n - p - 1$ degrés de liberté. On constate immédiatement qu'il s'agit du même test F que celui obtenu par analyse de la variance. Il suffit de remplacer dans l'équation (5.32) $\hat{R}$ par son expression (5.31).

On rejette H_0 si $F \geq Q_F(1 - \alpha; p, n - p - 1)$, sinon on ne rejette pas H_0.

5.4.3 Remarque

Considérons un vecteur $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ observé sur un échantillon d'effectif n. Si on isole la variable X_1 du vecteur et qu'on la considère comme la variable Y, on peut calculer la régression multiple de X_1 sur le vecteur à $p - 1$ dimensions $(X_2, \ldots, X_p)$ et le coefficient de corrélation multiple correspondant $\hat{R}_1$ comme décrit ci-dessus. On peut répéter la même opération pour chacune des variables du vecteur $\underset{\sim}{X}$. Au total, on peut donc calculer p régressions multiples et p coefficients de corrélation multiple $\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_p$. Il n'y a pas réellement de relation entre eux. Il ne faut pas confondre ces coefficients de corrélation multiple et la matrice $\underset{\sim}{R}$ des corrélations entre $X_1, \ldots, X_p$.

5.5 Méthodes de sélection de variables

En régression multiple, comme dans d'autres techniques de statistique multivariée, on s'efforce de ne retenir que les variables utiles et d'éliminer celles n'apportant qu'une information faible ou redondante. On parle de sélection de variables. Ce problème peut être approché de différentes manières.

Une première approche consiste à calculer la régression multiple de Y sur $\underset{\sim}{X}$, soit $\hat{Y} = b_0 + b_1 x_1 + \ldots + b_p x_p$, et de tester la signification statistique de chaque coefficient de régression en utilisant la méthode décrite à la section 5.3.4. On élimine les variables non utiles et on recommence les calculs de régression multiple sur les variables non éliminées.

On peut aussi faire autant de régression multiple qu'il y a de sous-ensembles de variables dans le vecteur $\underset{\sim}{X}$ et choisir celle qui donne les meilleurs résultats. Cette méthode est faisable lorsqu'on a peu de variables mais devient impossible lorsque p est grand, car le nombre de possibilités est proche de 2^p. Pour $p = 30$, on a plus d'un milliard de possibilités.

5.5.1 Sélection ascendante (forward)

La méthode de sélection ascendante pas à pas consiste à rechercher dans une première étape les régressions simples de Y sur chacune des variables $X_1, \ldots, X_p$. On choisit la meilleure, c'est-à-dire celle qui donne le coefficient de détermination $\hat{R}^2$ le plus élevé. Soit $X_{(1)}$ cette variable.

On calcule ensuite toutes les régressions multiples de Y sur $X_{(1)}$ et chacune des $p-1$ variables restantes. On retient celle qui donne le plus grand $\hat{R}^2$. Soit $X_{(2)}$ cette variable.

On calcule ensuite la régression multiple de Y sur tous les triplets ($X_{(1)}$, $X_{(2)}, X_i$) où X_i est chaque fois une des $p-2$ variables restantes. On procède ainsi de suite avec les quadruplets, quintuplets, etc.

On s'arrête à l'étape k lorsque plus aucune des $p-k$ variables restantes n'améliore de façon significative le $\hat{R}^2$ des k variables sélectionnées $X_{(1)}, \ldots, X_{(k)}$. Le critère utilisé est

$$F = \frac{\mathrm{SCR}_{(k+1)} - \mathrm{SCR}_{(k)}}{s^2_{(k+1)}}$$

qui est distribué comme un F de Snedecor à 1 et $n-k-2$ degrés de liberté. On arrête le processus dès que $F < Q_F(1-\alpha; 1, n-k-2)$ pour chacune des $p-k$ variables ajoutées.

5.5.2 Sélection descendante (backward)

La méthode procède par l'autre bout. On calcule d'abord la régression multiple de Y sur l'ensemble des variables $X_1, \ldots, X_p$. On calcule le $\hat{R}^2$ correspondant. Ensuite, dans une première étape, on retire tour à tour chacune des p variables. On obtient ainsi p régressions multiples basées sur $p-1$ variables. On calcule à chaque fois le $\hat{R}^2$. On élimine la variable qui donne le plus grand $\hat{R}^2$, c'est-à-dire celle qui fait perdre le minimum d'information. Soit $X_{(1)}$ cette variable. On calcule ensuite toutes les régressions multiples de Y en éliminant en plus de $X_{(1)}$ tour à tour chacune des $p-1$ variables restantes et on calcule le $\hat{R}^2$ à chaque fois. On élimine la variable $X_{(2)}$ qui donne le plus grand $\hat{R}^2$ et on continue ainsi de suite jusqu'à ce que l'élimination de chacune des variables restantes fasse chuter de façon significative $\hat{R}^2$.

On s'arrête à l'étape k lorsque chacune des $p-k$ variables restantes fait chuter significativement le $\hat{R}^2$ lorsqu'elle est éliminée. Le critère utilisé est

$$F = \frac{\mathrm{SCR}_{(k)} - \mathrm{SCR}_{(k-1)}}{s^2_{(k)}}$$

qui est distribué comme un F de Snedecor à 1 et $n-k-1$ degrés de liberté. On arrête le processus dès que $F \geq Q_F(1-\alpha; 1, n-k-1)$ pour chacune des k variables retenues.

5.5.3 Sélection "stepwise"

La méthode de sélection de variables dite "stepwise" combine à la fois la méthode ascendante et descendante. A chaque étape, on recherche le variable qui améliore au maximum le modèle mais avant de poursuivre on vérifie si, parmi les variables déjà sélectionnées, aucune ne doit être retirée.

5.6 Exemple

5.6.1 Données

A titre d'exemple, le taux de cholestérol (mg/dl), le poids (kg) et la pression artérielle systolique (PAS, mmHg) ont été mesurés chez 11 sujets masculins en bonne santé âgés de 14 à 24 ans.

Sujet	Cholestérol (mg/dl)	Poids (kg)	PAS (mmHg)
1	162.2	51.0	108
2	158.0	52.9	111
3	157.0	56.0	115
4	155.0	56.5	116
5	156.0	58.0	117
6	154.1	60.1	120
7	169.1	58.0	124
8	181.0	61.0	127
9	174.9	59.4	122
10	180.2	56.1	121
11	174.0	61.1	125

5.6.2 Régression multiple

Déterminons la régression multiple du taux de cholestérol (Y) sur le poids (X_1) et la PAS (X_2).
On calcule d'abord les sommes et sommes de carrés et produits croisés :

$$\begin{array}{lll} \sum y = 1821.5 & \sum x_1 = 630.1 & \sum x_2 = 1306 \\ \sum y^2 = 302723.51 & \sum yx_1 = 104467.79 & \sum yx_2 = 216682 \\ & \sum x_1^2 = 36197.45 & \sum x_1x_2 = 74983.3 \\ & & \sum x_2^2 = 155410 \end{array}$$

Notons que $\bar{y} = 165.59$
$\bar{x_1} = 57.282$
$\bar{x_2} = 118.73$
Calculons les éléments de la matrice $\underset{\sim}{A}_{2\times 2}$

$$a_{11} = \sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} = 36197.45 - \frac{(630.1)^2}{11} = 104.18$$

$$a_{12} = \sum x_1x_2 - \frac{(\sum x_1)(\sum x_2)}{n} = 74983.3 - \frac{630.1 \times 1306}{11} = 173.25$$

$$a_{22} = \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n} = 155410 - \frac{(1306)^2}{11} = 352.18$$

La matrice $\underset{\sim}{A}$ s'écrit donc

$$\underset{\sim}{A} = \begin{pmatrix} 104.18 & 173.25 \\ 173.25 & 352.18 \end{pmatrix}$$

De même, les éléments du vecteur $\underset{\sim}{c}$ valent respectivement

$$c_1 = \sum yx_1 - \frac{(\sum y)(\sum x_1)}{n} = 104467.79 - \frac{1821.5 \times 630.1}{11} = 128.96$$

$$c_2 = \sum yx_2 - \frac{(\sum y)(\sum x_2)}{n} = 216682 - \frac{1821.5 \times 1306}{11} = 420.27$$

de sorte que

$$\underset{\sim}{c} = \begin{pmatrix} 128.96 \\ 420.27 \end{pmatrix}.$$

Il faut alors calculer l'équation matricielle (5.11)

$$\begin{aligned} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} &= \begin{pmatrix} 104.18 & 173.25 \\ 173.25 & 325.18 \end{pmatrix}^{-1} \begin{pmatrix} 128.96 \\ 420.27 \end{pmatrix} \\ &= \begin{pmatrix} 0.0528 & -0.0260 \\ -0.0260 & 0.0156 \end{pmatrix} \begin{pmatrix} 128.96 \\ 420.27 \end{pmatrix} \end{aligned}$$

de sorte que $b_1 = -4.1039$ et $b_2 = 3.2121$.

Le terme indépendant vaut dès lors

$$\begin{aligned} b_0 &= \bar{y} - \underset{\sim}{b}^T \underset{\sim}{\bar{x}} \\ &= 19.30 \end{aligned}$$

La droite de régression a donc pour équation

$$\hat{Y} = 19.30 - 4.10\ x_1 + 3.21\ x_2$$

ou encore

$$\text{Cholestérol} = 19.30 - 4.10 \times \text{poids} + 3.21 \times \text{PAS}$$

5.6.3 Analyse de la variance

$$SCT = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 302723.51 - \frac{(1821.5)^2}{11} = 1099.67$$

$$SCR = \underset{\sim}{b}^T \underset{\sim}{c} = -4.1039 \times 128.96 + 3.2121 \times 420.27 = 820.71$$

$$SCE = SCT - SCR = 1099.67 - 820.71 = 278.96$$

Le tableau d'analyse de la variance s'établit comme suit

Variabilité	Somme de carrés	Degrés de liberté	Carré moyen	F
Régression	820.71	2	410.36	11.77
Résiduelle	278.96	8	s^2=34.866	
Totale	1099.67	10		

La valeur du F de Snedecor à 2 et 8 degrés de liberté est égale à 11.77. Comme le quantile à 95% vaut $Q_F(0.95; 2, 8) = 4.46$ et que $F = 11.77 > 4.46$, on rejette l'hypothèse $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$ et on conclut que la régression multiple du taux de cholestérol sur la poids et la PAS a un sens.

5.6.4 Utilité des variables explicatives

En utilisant l'équation (5.23) et en notant que $s = \sqrt{34.866} = 5.9047$, les erreurs types des coefficients de régression valent respectivement

$$\begin{aligned} SE(b_1) &= 5.9047 \times \sqrt{0.0528} = 1.3568 \\ SE(b_2) &= 5.9047 \times \sqrt{0.0156} = 0.7375 \end{aligned}$$

En comparant les valeurs absolues du t de Student au seuil critique bilatéral à 5%, soit $Q_t(0.975; 8) = 2.31$, on constate que les deux coefficients de régression sont utiles. En effet, les résultats du t de Student

$$\begin{aligned} \text{Poids :} \quad & t = \frac{b_1}{SE(b_1)} = \frac{-4.1039}{1.3568} = -3.02 \ (p = 0.016) \\ \text{PAS :} \quad & t = \frac{b_2}{SE(b_2)} = \frac{3.2121}{0.7375} = 4.36 \ (p = 0.0024) \end{aligned}$$

sont tous deux supérieurs à 2.31 en valeur absolue.

5.6.5 Qualité de la régression

En vertu de l'équation (5.25), on a

$$R^2 = \frac{SCR}{SCT} = \frac{820.71}{1099.67} = 0.746$$

En conséquence, 74.6% de la variabilité du taux de cholestérol sont expliqués par le poids et la PAS. Il reste 25.4% d'inexpliqué.

5.6.6 Prédiction et intervalle de confiance

Considérons un sujet dont le poids est égal à 55 kg et la PAS à 120 mmHg. La valeur prédite du taux de cholestérol vaut

$$\begin{aligned} \text{Cholestérol} &= 19.30 - 4.10 \times 55 + 3.21 \times 120 \\ &= 179.0 \text{ mg/dl} \end{aligned}$$

La variabilité d'échantillonnage de cette prédiction est donnée par l'équation (5.26),

$$\text{Var } \hat{Y} = 34.866 \left\{ \frac{1}{11} + h^2 \right\}$$

où

$$\begin{aligned} h^2 &= (55 - 57.28, 120 - 118.73) \begin{pmatrix} 0.0528 & -0.0260 \\ -0.0260 & 0.0156 \end{pmatrix} \begin{pmatrix} 55 - 57.28 \\ 120 - 118.75 \end{pmatrix} \\ &= (-2.28, 1.27) \begin{pmatrix} 0.0528 & -0.0260 \\ -0.0260 & 0.0156 \end{pmatrix} \begin{pmatrix} -2.28 \\ 1.27 \end{pmatrix} \\ &= 0.4497 \end{aligned} \tag{5.33}$$

Dès lors, Var $\hat{Y} = 18.849$ et SE($\hat{Y}$)=4.3415. On peut donc conclure que pour les personnes de poids égal à 55 kg et de pression artérielle systolique égale à 120 mmHg, le cholestérol moyen (mg/dl) est compris dans l'intervalle de confiance à 95%, $\hat{Y} \pm 1.96 SE(\hat{Y}) = 179.03 \pm 1.96 \times 4.3415 = [170.5 - 187.5]$.

5.6.7 Corrélation multiple

La corrélation multiple du cholestérol sur le poids et la PAS vaut

$$\hat{R}_1 = \sqrt{0.746} = 0.86$$

On pourrait de même calculer le coefficient de corrélation multiple du poids sur le cholestérol et la PAS ($\hat{R}_2 = \sqrt{0.92} = 0.96$) et celui de la PAS sur le poids et le cholestérol ($\hat{R}_3 = \sqrt{0.95} = 0.97$) puisqu'aucune variable n'est réellement privilégiée.

Chapitre 6

Régression logistique

6.1 Introduction

La régression logistique fait partie des méthodes de régression sur plusieurs variables. Elle s'applique lorsque la variable "dépendante" est binaire (et non plus continue). Cette situation est fréquente. Par exemple, on souhaite prédire l'issue d'un patient (vie ou décès, rémission ou non rémission) à partir d'un ensemble de covariables (variables "indépendantes") anamnestiques, cliniques, biochimiques. Donc, on s'efforce d'évaluer l'issue du patient avant que celle-ci ne se produise réellement. Peut-on prédire la réussite ou l'échec d'un étudiant avant qu'il n'entame des études supérieures ? Peut-on évaluer les chances de réussite d'un projet urbain financé sur base de caractéristiques de la ville qui sollicite le financement du projet ? On introduira le modèle de régression logistique et l'estimation des paramètres du modèle par la méthode du maximum de vraisemblance. Les techniques de sélection de variables seront brièvement évoquées. On terminera le chapitre en présentant la régression logistique ordinale qui généralise la méthode de régression logistique simple. Des exemples illustreront la méthodologie.

6.2 Définition du modèle

Soient Y une variable "dépendante" binaire (0 ou 1) et $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ un vecteur de variables "indépendantes" appelées aussi "covariables". Aucune restriction n'est faite sur le nombre et la nature des variables $X_1, \ldots, X_p$. Le problème posé consiste à trouver une relation entre la moyenne de Y et le vecteur $\underset{\sim}{X}$ comme on l'a fait en régression multiple.

Pour rappel, la moyenne d'une variable binaire $E(Y)$ est une proportion,

$$E(Y) = \pi \tag{6.1}$$

c'est-à-dire la probabilité que Y soit égal à 1, $P[Y = 1]$. De même, pour toute valeur $\underset{\sim}{X} = \underset{\sim}{x}$, on a

$$\begin{aligned} E(Y|\underset{\sim}{x}) &= P\left[Y = 1|\underset{\sim}{X} = \underset{\sim}{x}\right] \\ &= \pi(\underset{\sim}{x}) \end{aligned} \tag{6.2}$$

Par ailleurs, comme une proportion est toujours comprise entre 0 et 1, on peut écrire

$$0 \leq \pi(\underset{\sim}{x}) \leq 1 \quad \forall \underset{\sim}{x} \tag{6.3}$$

Dans la recherche d'une relation entre la moyenne Y et le vecteur $\underset{\sim}{X}$, l'approche utilisée en régression multiple (voir (5.2)), c'est-à-dire le modèle $E(Y|\underset{\sim}{x}) = \pi(\underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, n'est pas applicable car le membre de gauche est une proportion $\pi(\underset{\sim}{x})$ confinée dans l'intervalle [0, 1] en vertu de (6.3), alors que le membre de gauche $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}$ varie en principe de $-\infty$ à $+\infty$. Il faut donc trouver un autre modèle. L'idée est de transformer le membre de gauche $\pi(\underset{\sim}{x})$ de manière à le libérer de l'intervalle [0, 1] et de le faire varier dans l'intervalle $]-\infty, +\infty[$. Cette transformation donne lieu au modèle logistique.

Appliquons à $\pi(\underset{\sim}{x})$ la transformation *logit*

$$\text{logit } \pi(\underset{\sim}{x}) = \log \frac{\pi(\underset{\sim}{x})}{1 - \pi(\underset{\sim}{x})} \tag{6.4}$$

où "log" est le logarithme Népérien que l'on note souvent "ln". Dans ces conditions, le modèle de *régression logistique* s'écrit

$$\log \frac{\pi(\underset{\sim}{x})}{1 - \pi(\underset{\sim}{x})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{6.5}$$

ce qui ne pose plus de problème. Remarquons aussi que $1 - \pi(\underset{\sim}{x}) = P\left[Y = 0|\underset{\sim}{x}\right]$.

Le modèle (6.5) peut aussi s'écrire, en prenant l'exponentielle des deux membres et en solutionnant par rapport à $\pi(\underset{\sim}{x})$,

$$\pi(\underset{\sim}{x}) = \frac{e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}}} \tag{6.6}$$

où $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.

La terminologie de régression "logistique" provient du fait que la fonction $f(t) = e^t/(1+e^t), t \in \mathbb{R}$, est appelée la fonction logistique en mathématique. Dans l'équation (6.6), on constate qu'à la fois le membre de gauche et le membre de droite sont toujours compris dans l'intervalle [0, 1]. Le modèle (6.6), qui relie la moyenne de Y au vecteur $\underset{\sim}{X}$, n'est plus linéaire mais est une fonction logistique d'une combinaison linéaire des variables $X_1, \ldots, X_p$. Il définit la "régression logistique de Y sur $\underset{\sim}{X}$".

Notons que le modèle logistique peut aussi s'écrire

$$\pi(\underset{\sim}{x}) = \frac{1}{1 + e^{-(\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x})}} \tag{6.7}$$

On constate enfin que si $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} = 0, \pi(\underset{\sim}{x}) = P[Y = 1 | \underset{\sim}{x}] = 0.5$.

6.3 Estimateurs du maximum de vraisemblance

On part de la matrice d'observations

$$\underset{\sim}{X}_{n \times (p+1)} = \begin{pmatrix} y_1 & x_{11} & \ldots & x_{1p} \\ y_2 & y_{21} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ y_n & x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{6.8}$$

où les observations $y_i = 0$ ou 1 pour $\forall i$.

Comme en régression multiple, on peut distinguer deux situations, selon que les variables $X_1, \ldots, X_p$ ont été fixées et la variable Y observée, ou que la variable Y et le vecteur $\underset{\sim}{X}$ sont observés simultanément. Adoptons cette seconde approche.

Pour estimer les paramètres (coefficients) inconnus $\beta_0, \beta_1, \ldots, \beta_p$, la méthode des moindres carrés n'est plus applicable car la variable Y n'est plus Normale mais binaire. On a alors recours à la méthode du maximum de vraisemblance (ML = maximum likelihood).

Par définition, la *vraisemblance* de l'observation $\left(y_i, \underset{\sim}{x}_i^T\right)$, c'est-à-dire la ligne i de la matrice (6.8), s'écrit successivement

$$\begin{aligned} L(y_i, \underset{\sim}{x}_i^T) &= L(y_i | \underset{\sim}{x}_i) \cdot L(\underset{\sim}{x}_i) \\ &= \left[\frac{e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i}}{1 + e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i}}\right]^{y_i} \left[\frac{1}{1 + e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i}}\right]^{1-y_i} \cdot L(\underset{\sim}{x}_i) \end{aligned} \tag{6.9}$$

où $L(\underset{\sim}{x}_i)$ est la vraisemblance marginale du vecteur $\underset{\sim}{x}_i$ qui ne dépend pas des paramètres inconnus β_0 et $\underset{\sim}{\beta}$.

La vraisemblance totale de l'échantillon (6.8) est le produit des vraisemblances des observations $\left(y_i, \underset{\sim}{x}_i^T\right)$ et on a

$$\begin{aligned} L\left(\beta_0, \underset{\sim}{\beta}^T\right) &= \prod_{i=1}^{n} L(y_i, \underset{\sim}{x}_i) \\ &= \prod_{i=1}^{n} L(y_i|\underset{\sim}{x}_i) \cdot L(\underset{\sim}{x}_i) \end{aligned} \tag{6.10}$$

En prenant le logarithme des deux membres de l'équation (6.10) et en tenant compte de l'expression (6.9), on obtient le logarithme de la vraisemblance soit

$$\begin{aligned} & l\left(\beta_0, \underset{\sim}{\beta}^T\right) \\ &= \log L\left(\beta_0, \underset{\sim}{\beta}^T\right) \qquad (6.11) \\ &= \sum_{i=1}^{n}\left[y_i \log \frac{e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}} + (1-y_i)\log\frac{1}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}\right] + \sum_{i=1}^{n}\log L(\underset{\sim}{x}_i) \end{aligned}$$

Pour une matrice d'observations (6.8) donnée, la fonction $l\left(\beta_0, \underset{\sim}{\beta}^T\right)$ ne dépend que des paramètres inconnus. On peut même laisser tomber le second terme de (6.11) car il ne dépend pas de $\underset{\sim}{\beta}$ et on a

$$l\left(\beta_0, \underset{\sim}{\beta}^T\right) = \sum_{i=1}^{n}\left[y_i \log \frac{e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}} + (1-y_i)\log\frac{1}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}\right] + C \tag{6.12}$$

Les estimateurs (ML) du maximum de vraisemblance $\hat{\beta}_0$ et $\underset{\sim}{\hat{\beta}}$ (ou $b_0, \underset{\sim}{b}$) sont ceux qui maximisent la fonction (6.12). Il faut donc maximiser une fonction à $p+1$ paramètres inconnus. On ne peut résoudre ce problème de façon "analytique" (c'est-à-dire par une formule comme en régression multiple) mais par une approche numérique, par exemple par la méthode de Newton-Raphson. Ceci n'est possible que par ordinateur.

L'ordinateur fournit les solutions $b_0, b_1, \ldots, b_p$ ainsi que les erreurs types $SE(b_0)$, $SE(b_1), \ldots, SE(b_p)$. Plus généralement, l'ordinateur fournit également la matrice de dispersion des estimateurs b_0 et $\underset{\sim}{b}$. On peut alors écrire

l'équation de prédiction

$$\hat{\pi}(\underset{\sim}{x}) = \hat{P}[Y = 1|\underset{\sim}{x}] = \frac{e^{b_0+b_1x_1+\ldots+b_px_p}}{1 + e^{b_0+b_1x_1+\ldots+b_px_p}} \tag{6.13}$$

et l'appliquer en pratique.
Note : Lorsque l'échantillon (6.8) est obtenu conditionnellement à $\underset{\sim}{X} = \underset{\sim}{x}$ fixé par l'utilisateur, on retrouve la même fonction de vraisemblance (6.12) et donc les mêmes estimateurs que si $\underset{\sim}{X}$ avait été observé en même temps que Y.

6.4 Tests sur le modèle

6.4.1 Approche globale

Comme en régression multiple, on peut s'interroger sur l'utilité du vecteur $\underset{\sim}{X}$ en testant l'hypothèse nulle globale

$$H_0 : \underset{\sim}{\beta} = \underset{\sim}{0} \text{ versus } H_1 : \underset{\sim}{\beta} \neq \underset{\sim}{0} \tag{6.14}$$

A cet effet, on calcule le critère du rapport de vraisemblance (likelihood ratio test)

$$LR = 2(\hat{l}_p - \hat{l}_0), \tag{6.15}$$

où $\hat{l}_p$ représente le maximum de vraisemblance basé sur l'ensemble des variables et $\hat{l}_0$ celui obtenu en n'incluant que le terme indépendant β_0, qui est distribué comme une loi chi-carré à p degrés de liberté. Si $LR \geq Q_{\chi^2}(1-\alpha; p)$, on rejette H_0 et la régression logistique a un sens. Dans le cas contraire, on peut supposer $\underset{\sim}{\beta} = \underset{\sim}{0}$. Dès lors, il n'y pas d'association entre Y et $\underset{\sim}{X}$.

6.4.2 Utilité des covariables

Lorsque le vecteur $\underset{\sim}{\beta}$ n'est pas nul, on peut s'interroger sur l'utilité de chacune des variables $X_1, \ldots, X_p$ dans le modèle en testant les hypothèses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ pour chaque variable $(i = 1, \ldots, p)$.

A cet effet, on calcule le critère

$$Z_i = \frac{b_i}{\mathrm{SE}(b_i)} \tag{6.16}$$

ou son carré Z_i^2, distribué comme un chi-carré à 1 degré de liberté. Donc, on rejette l'hypothèse $\beta_i = 0$ si

$$Z_i^2 \geq Q_{\chi^2}(1 - \alpha; 1) \tag{6.17}$$

Comme en général $\alpha = 0.05$, le seuil de décision est égal à $Q_{\chi^2}(0.95; 1) = 3.84$.

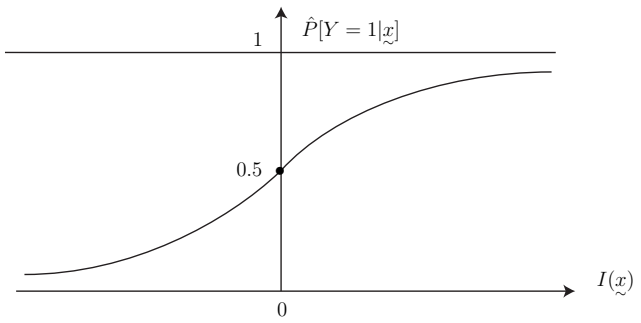
6.4.3 Qualité du modèle logistique

Pour apprécier la qualité du modèle logistique, on envisage toutes les paires d'observations (y_i, y_j), où $y_i = 0$ et $y_j = 1$. Si on note n_0 le nombre d'observations pour lesquelles $y_i = 0$ et n_1 le nombre d'observations pour lesquelles $y_j = 1$, on constate qu'il y a $N = n_0 \times n_1$ paires distinctes.

On calcule ensuite pour chaque paire (y_i, y_j), les probabilités $\hat{\pi}(\underset{\sim}{x}_i)$ et $\hat{\pi}(\underset{\sim}{x}_j)$ à partir de la formule (6.13). Puisque $y_i < y_j$, on devrait avoir $\hat{\pi}(\underset{\sim}{x}_i) < \hat{\pi}(\underset{\sim}{x}_j)$. Si tel est le cas, on dit qu'il y a *concordance*. Si tel n'est pas le cas, on dit qu'il y a *discordance*. Lorsque $\hat{\pi}(\underset{\sim}{x}_i) = \hat{\pi}(\underset{\sim}{x}_j)$, on dit qu'il y a *ex-aequo* ("tied pairs", en anglais). Désignons par N_c le nombre de paires "concordantes". Dans ce cas, la qualité de la régression est appréciée par le pourcentage de concordance, soit $(N_c/N) \times 100\%$. On calcule aussi les pourcentages de discordance et d'ex-aequo.

6.4.4 Prédiction

L'équation (6.13) permet de prédire pour toute nouvelle observation $\underset{\sim}{x}$ la probabilité que $Y = 1$, c'est-à-dire la proportion de sujets ayant la valeur $Y = 1$ parmi tous les sujets ayant la même caractéristique $\underset{\sim}{x}$. On constate qu'on peut prédire cette probabilité au travers du scalaire $I(\underset{\sim}{x}) = b_0 + \underset{\sim}{b}^T \underset{\sim}{x}$, somme pondérée des valeurs du vecteur $\underset{\sim}{X}$. On peut reporter sur un graphique $\hat{P}[Y = 1|\underset{\sim}{x}]$ en fonction de l'index $I(\underset{\sim}{x})$ et on retrouve une courbe logistique.



En clair, à partir de l'observation $\underset{\sim}{X} = \underset{\sim}{x}$, on calcule l'index de risque $I(\underset{\sim}{x})$ et en utilisant l'équation (6.13) qui s'écrit aussi

$$\hat{P}\left[Y = 1|\underset{\sim}{x}\right] = \frac{e^{I(\underset{\sim}{x})}}{1 + e^{I(\underset{\sim}{x})}} \tag{6.18}$$

on obtient la probabilité de l'événement $Y = 1$. On parle souvent de $I(\underset{\sim}{x})$ comme d'un index de risque.

Comme en régression multiple, on peut calculer un intervalle de confiance à 95% pour $\hat{P}[Y = 1|\underset{\sim}{x}]$. On calcule d'abord un intervalle de confiance pour $I(\underset{\sim}{x})$, dont les limites sont $I_1(\underset{\sim}{x})$ et $I_2(\underset{\sim}{x})$. A cet effet, on a besoin de la matrice de variances - covariances estimée des estimations b_0 et $\underset{\sim}{b}$, notée $\underset{\sim}{\hat{V}}$. Celle-ci est obtenue par ordinateur. Ensuite, en utilisant l'équation (6.6) et en y substituant à droite $I_1(\underset{\sim}{x})$ puis $I_2(\underset{\sim}{x})$ on obtient une fourchette pour la probabilité.

6.5 Méthodes de sélection de variables

En régression logistique, on est également confronté au problème de sélection de variables. On souhaite ne retenir que les variables utiles. Plusieurs approches sont possibles. On peut se contenter de ne retenir que les variables pour lesquelles $\beta_i \neq 0$ (voir section 6.4.2.). On peut aussi considérer tous les sous-ensembles de variables et faire à chaque fois une régression logistique. Cette façon de procéder est quasiment impossible tant le nombre de possibilités est élevé. En outre, le temps de calcul ordinateur s'accroît de façon vertigineuse. On se tourne comme en régression multiple vers des méthodes ascendante, descendante ou "stepwise".

6.5.1 Sélection ascendante (forward)

On réalise dans une première étape les régressions logistiques de Y sur chacune des variables X_i prises séparément. On choisit la meilleure, c'est-à-dire celle qui donne la plus grande vraisemblance maximisée (6.12). Soit $X_{(1)}$ cette variable et $\hat{l}_1 = \hat{l}(b_0, b_1)$ le maximum de vraisemblance.

On calcule ensuite toutes les régressions logistiques à deux variables, en incluant $X_{(1)}$ et chacune des variables restantes tout à tour. On retient la variable $X_{(2)}$ qui avec $X_{(1)}$ donne la plus grande vraisemblance estimée, soit $\hat{l}_2 = \hat{l}(b_0, b_1, b_2)$. On calcule ensuite les régressions logistiques de Y sur tous les

triplets $(X_{(1)}, X_{(2)}, X_i)$ où X_i est chaque fois une des $p-2$ variables restantes. On procède ainsi de suite avec les quadruplets, quintruplets, sextuplets, etc.

On s'arrête à l'étape k lorsque plus aucune des $p-k$ variables restantes n'améliore de façon significative le maximum de vraisemblance des k variables sélectionnées. Le critère utilisé s'écrit

$$\chi^2_{(1)} = 2\left(\hat{l}_{k+1} - \hat{l}_k\right) \tag{6.19}$$

distribué comme un chi-carré à 1 degré de liberté. Dès lors, on s'arrête à l'étape k si $\chi^2_{(1)} < 3.84$ pour chacune des $p-k$ variables restantes.

6.5.2 Sélection descendante (backward)

On procède comme en régression multiple en partant du modèle logistique complet à p variables. On calcule le maximum de vraisemblance $\hat{l}_p$. Ensuite, on retire parmi les variables $X_1, \ldots, X_p$ celle pour laquelle le maximum de vraisemblance, $\hat{l}_{p-1}$, diminue le moins par rapport à $\hat{l}_p$. Soit $X_{(1)}$ cette variable. On procède de même à partir des $p-1$ variables restantes et on recherche celle qui fait chuter le moins le maximum de vraisemblance $\hat{l}_{p-2}$. Soit $X_{(2)}$ cette variable. On continue de la sorte jusqu'à l'arrêt du processus.

On s'arrête à l'étape k du processus lorsque chacune des $p-k$ variables restantes détériore significativement le maximum de vraisemblance $\hat{l}_{p-k}$ lorsqu'elles sont retirées du problème. En d'autres termes, on arrête le processus à l'étape k si

$$\chi^2_{(1)} = 2\left(\hat{l}_k - \hat{l}_{k-1}\right) \tag{6.20}$$

est statistiquement significatif ($\chi^2_{(1)} \geq 3.84$) chaque fois qu'on retire une variable des $p-k$ toujours en lice.

6.5.3 Sélection "stepwise"

La méthode de sélection de variables dite "stepwise" combine à la fois la méthode ascendante et descendante. A chaque étape, on recherche le variable qui améliore au maximum le modèle mais avant de poursuivre on vérifie si, parmi les variables déjà sélectionnées, aucune ne doit être retirée.

6.6 Odds ratio

En épidémiologie, l'*odds ratio* (OR) est une quantité fréquemment utilisée pour mesurer l'association entre un facteur de risque (variable X) et la

présence ou non d'une maladie (variable Y). L'odds ratio relatif à la variable $X_i (i = 1, \ldots, p)$ et ajusté pour les autres variables est donné par l'expression

$$OR_i = e^{\beta_i} \tag{6.21}$$

On voit que si $\beta_i = 0$, il n'y a pas d'association entre le facteur de risque X_i et la maladie ($OR_i = 1$). Par contre, si $\beta_i > 0$ ou $\beta_i < 0$, il y a une association positive ($OR_i > 1$) ou négative ($OR_i < 1$) entre les deux.

On peut estimer OR_i grâce à la méthode de régression logistique et on a immédiatement

$$\widehat{OR}_i = e^{b_i} \tag{6.22}$$

estimation qui s'interprète de la même manière que la valeur théorique.

Puisqu'on connaît $SE(b_i)$, on peut déterminer un intervalle de confiance à 95% pour OR_i en calculant les quantités $b_{i1} = b_i - 1.96 \times SE(b_i)$ et $b_{i2} = b_i + 1.96 \times SE(b_i)$ et ensuite $\widehat{OR}_{i1} = e^{b_{i1}}$ et $\widehat{OR}_{i2} = e^{b_{i2}}$. L'intervalle de confiance à 95% (IC 95%) s'écrit alors

$$\widehat{OR}_{i1} \leq OR_i \leq \widehat{OR}_{i2} \quad (i = 1, \ldots, p) \tag{6.23}$$

Si cet intervalle de confiance contient la valeur 1, il n'y a pas d'association entre le facteur de risque X_i et la maladie Y. Dans le cas contraire, l'association est significative. En d'autres termes, il suffit de regarder si l'intervalle de confiance contient ou non la valeur 1.

6.7 Exemple

Les données reprises à l'Annexe II représentent chez 60 traumatisés crâniens, l'issue à six mois (Y) en trois catégories (1 = bonne récupération ou invalidité légère, 2 = invalidité sévère ou état végétatif persistant, 3 = décès) ainsi que l'âge au moment de l'accident et le taux de CK-BB (isoenzyme cérébrale de la créatine kinase, UI/l) dans le liquide céphalo-rachidien mesuré dans les 24 heures après l'accident. Pour illustrer la méthode de régression logistique, on regroupe les catégories 1 et 2 de l'issue pour former un variable binaire $Y = 0$ (sujet en vie) et $Y = 1$ (sujet décédé). Il y a 27 sujets (45%) en vie et 33 sujets (55%) décédés.

L'application de la méthode de régression logistique conduit aux résultats suivants :

Paramètre	Estimation MV	Erreur type	Z^2	p-value
Intercept	$b_0 = -2.419$	0.7467	10.5	0.0012
Age	$b_1 = 0.0502$	0.0228	4.85	0.028
CK-BB	$b_2 = 0.00534$	0.00181	8.72	0.0032

L'index de risque s'écrit donc

$$I(\underset{\sim}{x}) = -2.42 + 0.0502 \times \text{âge} + 0.00534 \times \text{CK-BB}$$

On constate que les deux covariables, âge et CK-BB, sont significatives pour prédire l'issue du patient. Plus le patient est âgé et plus le taux de CK-BB est élevé, plus le risque de décéder est accru. D'ailleurs, le test du rapport de vraisemblance montre que le vecteur $\underset{\sim}{\beta}$ n'est pas nul. En effet, on trouve $LR = 23.33$ à 2 degrés de liberté ($p < 0.0001$).

Pour mesurer la qualité du modèle, on observe qu'il y a au total $27 \times 33 = 891$ paires d'observations des deux sous-groupes de patients. Parmi celles-ci, 736 sont "concordantes" (82.6%), 152 sont "discordantes" (17.1%) et il y a 3 ex-aequos (0.3%). On peut dès lors conclure que le modèle est satisfaisant.

Considérons un patient âgé de 45 ans et présentant un taux de CK-BB de 300 UI/l. L'index de risque vaut

$$\begin{aligned} I(\underset{\sim}{x}) &= -2.42 + 0.0502 \times 45 + 0.00534 \times 300 \\ &= 1.44 \end{aligned}$$

La probabilité de décès correspondante a pour valeur

$$P[Y = 1|\underset{\sim}{x}] = \frac{e^{1.44}}{1 + e^{1.44}} = 0.81$$

On peut donc conclure que ce patient a 81% de chance de décéder et 19% de chance de survivre à son traumatisme crânien.

Notons que les odd ratios relatifs aux variables âge et CK-BB valent respectivement 1.05 et 1.01. Les intervalles de confiance correspondants s'écrivent $1.0055 < OR_1 < 1.0995$ et $1.0018 < OR_2 < 1.0089$. Tous deux ne contiennent pas la valeur $OR = 1$, confirmant le caractère significatif de l'association des deux variables avec l'issue du patient.

6.8 Régression logistique ordinale

6.8.1 Définition du problème

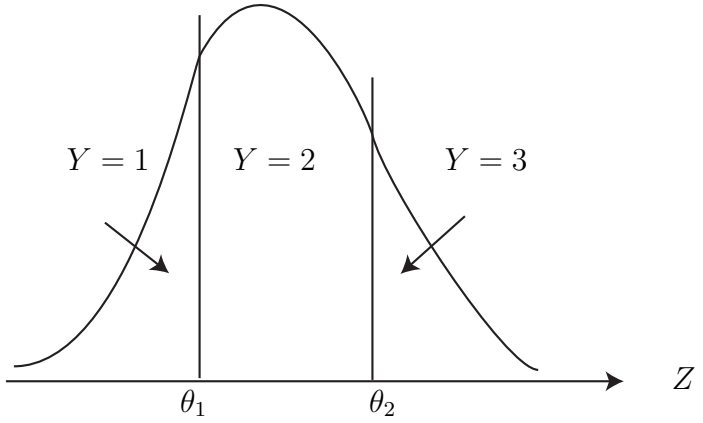
Il arrive fréquemment que la variable dépendante Y soit qualitative ordinale ; en d'autres termes, ses modalités (appelées aussi "catégories") sont ordonnées $m_1 < m_2 < \ldots < m_k$ (par exemple, les grades aux examens). Nous

avons vu qu'on pouvait numéroter les catégories de 1 à k. Sous la variable Y se cache en réalité une variable continue Z et des valeurs "seuils" $\theta_1, \ldots, \theta_{k-1}$ sur l'échelle de cette variable qui définissent la variable Y. Ainsi,

$$Y = i \text{ si } \theta_{i-1} < Z \leq \theta_i \quad (i = 1, \ldots, k) \tag{6.24}$$

Par définition, $\theta_0 = -\infty$ et $\theta_k = +\infty$. C'est un peu ce qui se passe lorsqu'en délibération on attribue un grade sur base de la moyenne des examens de l'étudiant. Prenons l'exemple de 3 catégories ordonnées $m_1 < m_2 < m_3$, on aurait

$$\begin{array}{lll} Y = 1 & \text{si} & Z \leq \theta_1 \\ Y = 2 & \text{si} & \theta_1 < Z \leq \theta_2 \\ Y = 3 & \text{si} & \theta_2 < Z \end{array} \tag{6.25}$$



En pratique, on ne connaît pas Z et on observe uniquement les catégories au travers de la variable Y. Cette approche s'applique aussi au modèle logistique. Dans ce cas, il n'y a que $k = 2$ catégories et donc un seul seuil (cut-off) θ. Dès lors, $Y = 0$ si $Z \leq \theta$ et $Y = 1$ si $Z > \theta$.

6.8.2 Modèle logistique ordinal

Supposons que l'on observe simultanément $(Y, \underset{\sim}{X})$ où Y est une variable ordinale à k modalités. On calcule la probabilité d'occurence de chaque modalité $P[Y = i] = \pi_i$ $(i = 1, \ldots, k)$. Bien sûr, $\pi_1 + \pi_2 + \ldots + \pi_k = 1$. De même, on peut travailler conditionnellement à $\underset{\sim}{X} = \underset{\sim}{x}$ et on a

$$\begin{array}{l} P\left[Y = i | \underset{\sim}{x}\right] = \pi_i(\underset{\sim}{x}) \\ \pi_1(\underset{\sim}{x}) + \pi_2(\underset{\sim}{x}) + \ldots + \pi_k(\underset{\sim}{x}) = 1 \; \forall \underset{\sim}{x} \end{array} \tag{6.26}$$

Les modalités de Y sont définies par les seuils $\theta_1, \ldots, \theta_{k-1}$ sur l'échelle de la variable cachée (on dit aussi "latente") Z.

On postule le modèle logistique ordinal suivant ($k = 3$ pour des raisons de clarté).

$$\begin{aligned} P\left[Y = 1|\underset{\sim}{x}\right] &= \frac{e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \\ P\left[Y = 2|\underset{\sim}{x}\right] &= \frac{e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} - \frac{e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \\ P\left[Y = 3|\underset{\sim}{x}\right] &= 1 - \frac{e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \end{aligned} \tag{6.27}$$

ce qui s'écrit pour le cas général

$$P\left[Y = i|\underset{\sim}{x}\right] = \frac{e^{\theta_i - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_i - \underset{\sim}{\beta}^T \underset{\sim}{x}}} - \frac{e^{\theta_{i-1} - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_{i-1} - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \qquad (i = 1, \ldots, k) \tag{6.28}$$

On constate que dans (6.27) et (6.28) la somme des probabilités est égale à 1. On reconnaît à chaque fois la fonction logistique $f(t) = e^t/(1 + e^t)$. Rappelons que $\underset{\sim}{\beta}^T \underset{\sim}{x} = \beta_1 x_1 + \ldots + \beta_p x_p$.

Les paramètres du modèle sont les $k - 1$ valeurs seuils $\theta_1, \ldots, \theta_{k-1}$ et les coefficients de régression des variables $\beta_1, \ldots, \beta_p$. Il n'y a pas de terme indépendant β_0, celui-ci étant remplacé par les valeurs seuils. Au total, on dénombre $p + k - 1$ paramètres différents qu'il faut estimer. Notons que si $k = 2$ (modèle logistique), on retrouve les $p + 1$ paramètres $\theta_1 = \beta_0$ et $\underset{\sim}{\beta}$.

6.8.3 Estimation du maximum de vraisemblance

La matrice d'observations se présente sous la même forme que celle donnée en (6.3), mais à présent les valeurs y_i sont égales à $1, 2, \ldots, k$.

Comme en régression logistique, on maximise la vraisemblance de l'échantillon ou son logarithme

$$l = \log L = \sum_{i=1}^{n} \log P\left[Y_i = y_i|\underset{\sim}{x}_i\right] + C \tag{6.29}$$

où on remplace les probabilités par leurs expressions (6.28). Pour y_i et $\underset{\sim}{x}_i$ donnés, cette vraisemblance ne dépend plus que des paramètres $\underset{\sim}{\theta}$ et $\underset{\sim}{\beta}$, soit

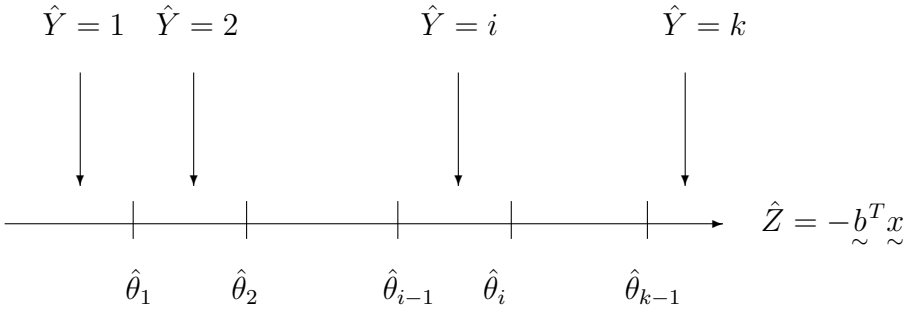
$l(\theta_1, \ldots, \theta_{k-1}, \beta_1, \ldots, \beta_p)$. Par ordinateur, on recherche les valeurs $\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}$, $b_1, \ldots, b_p$ qui maximisent (6.29) et on note $\hat{l}(\underset{\sim}{\hat{\theta}}, \underset{\sim}{b})$ le maximum obtenu. Le programme fournit aussi les erreurs types $\mathrm{SE}(\hat{\theta}_i)$ et $\mathrm{SE}(b_i)$ des estimations $\hat{\theta}_i$ et b_i. Les équations (6.28) deviennent

$$\hat{P}\left[Y = i | \underset{\sim}{x}\right] = \frac{e^{\hat{\theta}_i - \underset{\sim}{b}^T \underset{\sim}{x}}}{1 + e^{\hat{\theta}_i - \underset{\sim}{b}^T \underset{\sim}{x}}} - \frac{e^{\hat{\theta}_{i-1} - \underset{\sim}{b}^T \underset{\sim}{x}}}{1 + e^{\hat{\theta}_{i-1} - \underset{\sim}{b}^T \underset{\sim}{x}}} \qquad (i = 1, \ldots, k) \qquad (6.30)$$

6.8.4 Prédiction

Si on considère un nouveau sujet pour lequel on connaît $\underset{\sim}{x}$, on peut calculer les chances qu'il a d'avoir chacune des modalités de Y en utilisant les équations (6.30) et lui attribuer, par exemple, la catégorie pour laquelle la probabilité est la plus élevée.

On pourrait presque affirmer que l'index univarié $I(\underset{\sim}{x}) = -\underset{\sim}{b}^T \underset{\sim}{x}$ (attention on change le signe de tous les b_i !) est un prédicteur de la variable sous-jacente inobservable Z, une sorte d'indicateur de risque.



Il suffit de regarder où se situe la valeur $\hat{Z} = -\underset{\sim}{b}^T \underset{\sim}{x}$ par rapport aux différentes valeurs seuils $\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}$ et déterminer ainsi la catégorie Y.

6.8.5 Autres remarques

En régression logistique ordinale, comme en régression logistique simple, on peut apprécier la qualité de la régression en calculant le pourcentage de concordance pour toutes les paires (y_i, y_j) où $y_i < y_j$. Si on note, $n_1,\ n_2, ...,\ n_k$ le nombre observé de sujets dans chaque modalité de Y, il y a $N = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j$ paires. Une paire (y_i, y_j) conduit à des valeurs concordantes si $\hat{\pi}_i(\underset{\sim}{x}_i) < \hat{\pi}_j(\underset{\sim}{x}_j)$, à un ex-aequo s'il y a égalité et à une discordance

sinon. Les probabilités $\hat{\pi}_i(\underset{\sim i}{x})$ et $\hat{\pi}_j(\underset{\sim j}{x})$ sont calculées à partir de l'équation (6.13).

Les techniques de sélection de variables s'appliquent également, que ce soit de façon ascendante ou descendante.

Enfin, il est possible de tester l'égalité de deux seuils adjacents, par exemple $H_0 : \theta_i = \theta_{i+1}$. Si cette hypothèse n'est pas rejetée, on peut fusionner les deux seuils et donc les deux catégories et simplifier le modèle logistique (6.30). On dit dans ce cas que les deux catégories fusionnées sont indiscernables.

6.8.6 Exemple

Il est intéressant d'appliquer le modèle de régression logistique ordinale aux données des traumatisées crâniens. A présent, on considère un modèle à trois catégories définies par la variable Y. Notons "Evolution favorable" la catégorie $Y = 1$, "Evolution défavorable" la catégorie $Y = 2$ et "Décès" la catégorie 3. Il y a respectivement 19, 8 et 33 sujets dans les trois catégories ainsi définies.

Le modèle comporte au total 4 paramètres, soient les deux seuils θ_1 et θ_2 et les coefficients β_1 et β_2 des deux variables âge et CK-BB.

L'application du modèle de régression logistique ordinale conduit aux résultat suivants :

Paramètre	Estimation MV	Erreur type	Z^2	p-value
Seuil 1	$\hat{\theta_1} = 1.778$	0.6709	7.02	0.0081
Seuil 2	$\hat{\theta_2} = 2.629$	0.725	13.1	0.0003
Age	$b_1 = -0.0498$	0.0215	5.37	0.0204
CK-BB	$b_2 = -0.00601$	0.00188	10.2	0.0014

L'index de risque (variable latente Z) s'écrit

$$\begin{aligned} \hat{Z} &= -(-0.0498 \times \text{âge} - 0.00601 \times \text{CK-BB}) \\ &= 0.0498 \times \text{âge} + 0.00601 \times \text{CK-BB} \end{aligned}$$

que l'on peut confronter aux deux seuils de décision 1.78 et 2.63.

A nouveau, on constate que les deux variables âge et CK-BB sont significativement prédictives de l'issue Y. Le test du rapport de vraisemblance qui compare le modèle (complet) avec les deux covariables à celui qui ne contient que les deux seuils a pour valeur $LR = 27.46$ à 2 degrés de liberté ($p < 0.0001$).

La qualité du modèle est appréciée à partir du pourcentage de concordance obtenu sur les $N = 19 \times 8 + 19 \times 33 + 8 \times 33 = 1043$ paires possibles. On a trouvé 846 paires concordantes (81.1%), 194 paires discordantes (18.6%) et 3 paires ex-aequos (0.3%).

Reprenons l'exemple du patient âgé de 45 ans et présentant un taux de CK-BB de 300 UI/l. Son index de risque est égal à

$$\hat{Z} = 0.0498 \times 45 + 0.00601 \times 300 = 4.04$$

On constate qu'il se situe au-delà du second seuil puisque $4.04 > 2.63$. Le sujet tombe donc dans la catégorie "Décès". Les probabilités associées à chaque catégorie sont obtenues à partir des équations (6.27). On a respectivement

$$\begin{aligned}
\hat{\pi}_1(\underset{\sim}{x}) &= \frac{e^{1.78-4.04}}{1+e^{1.78-4.04}} = \frac{e^{-2.26}}{1+e^{-2.26}} = 0.09 \\
\hat{\pi}_2(\underset{\sim}{x}) &= \frac{e^{2.63-4.04}}{1+e^{2.63-4.04}} - \hat{\pi}_1(x) \\
&= \frac{e^{-1.41}}{1+e^{-1.41}} - 0.09 = 0.20 - 0.09 = 0.11 \\
\hat{\pi}_3(\underset{\sim}{x}) &= 1 - \hat{\pi}_1(\underset{\sim}{x}) - \hat{\pi}_2(\underset{\sim}{x}) \\
&= 1 - 0.09 - 0.11 = 0.80
\end{aligned}$$

On retrouve un résultat proche de celui obtenu par la régression logistique simple mais mieux différencié sur les deux autres catégories.

Chapitre 7

Régression de Cox

7.1 Introduction

La méthode de régression de Cox s'inscrit dans le domaine de l'analyse des durées de vie. L'étude des durées de survie occupe aujourd'hui une large place dans la littérature scientifique. Ce qui caractérise l'observation d'une variable de durée de vie par rapport aux autres variables, c'est la possibilité d'être censurée. Une observation censurée est une observation dont on ne connaît pas la valeur exacte mais dont on sait seulement qu'elle est supérieure (censure à droite) ou inférieure (censure à gauche) à une valeur donnée. Dans ce chapitre, on ne s'intéresse qu'aux censures à droite. On définira les notions de variable de durée de vie et de la courbe de survie qui lui est associée. On rappellera l'estimation de la courbe de survie par la méthode de Kaplan-Meier. On introduira ensuite le modèle des risques proportionnels de Cox et on s'intéressera à l'estimation des paramètres du modèle. Le problème de sélection de variables sera brièvement abordé et on clôturera par quelques applications du modèle.

7.2 Durée de vie

Une variable de *durée de vie*, notée T pour la circonstance, est une variable continue qui représente le temps écoulé entre deux événements. Par exemple, la durée de vie d'un être humain est l'intervalle de temps qui sépare sa naissance de son décès. C'est l'exemple le plus naturel d'une variable de durée de vie. Il peut aussi s'appliquer aux animaux, aux équipements, aux aliments, aux plantes, aux êtres vivants en général. Une durée de vie c'est aussi le temps d'hospitalisation d'un patient, la durée de chômage d'un ouvrier, la durée d'un mariage, le temps qui s'écoule entre deux crises d'asthme.

En pratique, une durée de vie s'exprime généralement en unité de temps (sec, min, heure, jour, mois, année).

Une durée de vie T se caractérise généralement par trois aspects :

1. $T \geq 0$ est une variable non négative
2. la distribution de T présente une forte dissymétrie à droite (donc la moyenne de T est supérieure à la médiane)
3. T est souvent censurée à droite (par exemple, $T > t$)

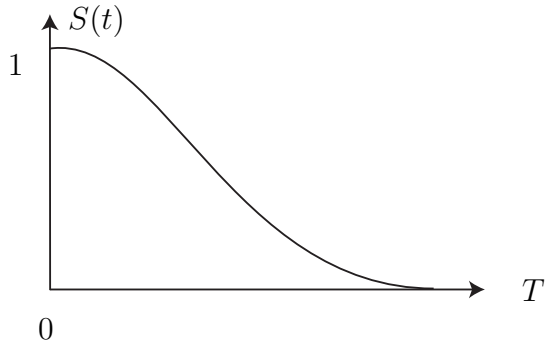
Les observations de la variable T sur un échantillon de n sujets (ou objets) doivent dès lors être séparées entre valeurs réelles (non censurées) et valeurs censurées. Il y a différentes manières d'indiquer qu'une donnée est censurée (par exemple, > 25 ou 25^* ou 25^+). Lorsqu'on encode les données dans un ordinateur en vue d'un traitement statistique, cette présentation n'est pas suffisante. A cet effet, on utilise un indicateur de censure C tel que $c = 0$ si l'observation est une durée exacte ($T = t$) et $c = 1$ si l'observation est censurée ($T > t$). En clair, on introduit la paire d'observations $(t_i, c_i), i = 1, \ldots, n$.

7.3 Courbe de survie de Kaplan-Meier

Une durée de vie T est caractérisée par sa *courbe de survie* ("survival curve")

$$S(t) = P[T > t] \qquad t \geq 0 \tag{7.1}$$

qui représente la proportion de sujets toujours en vie après t ou encore la probabilité de survie au-delà de t. On voit immédiatement que si $t = 0, S(0) = 1$. Par ailleurs, plus t augmente, plus $S(t)$ diminue. Ainsi, si $t_1 > t_2$, alors $S(t_1) \leq S(t_2)$. Enfin, lorsque $t \to \infty, S(t) \to 0$.



En pratique, pour estimer la courbe de survie à partir d'un échantillon de données $\{(t_i, c_i), i = 1, \ldots, n\}$, on utilise la *méthode de Kaplan-Meier*. Celle-ci procède comme suit

1. On trie les n données de survie t_i par ordre croissant. Dans ce classement, une durée censurée suit toujours une durée de vie non censurée de même valeur.
2. On ne retient que les k valeurs non censurées distinctes notées $t_1 < t_2 < \ldots < t_{k-1} < t_k$.
3. Pour chaque valeur t_i distincte, on note l_i le nombre de sujets toujours en vie juste avant ce moment et d_i le nombre de sujets qui décèdent en $T = t_i$. Attention, une durée censurée avant t_i n'est pas prise en compte car on ignore ce qu'elle est devenue.
4. On estime la valeur de la courbe de survie en $T = t_i$ par la formule

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{l_j - d_j}{l_j} \qquad (i = 1, \ldots, k) \tag{7.2}$$

ou encore par la formule de récurrence $(i = 1, \ldots, k)$

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \times \frac{l_i - d_i}{l_i} \tag{7.3}$$

où par convention si $i = 1, t_{i-1} = 0$ et $\hat{S}(0) = 1$.

Notons qu'en pratique la courbe de survie de Kaplan-Meier part toujours de 1 mais n'aboutit pas nécessairement à 0. C'est le cas lorsque la valeur la plus élevée de l'échantillon est censurée. On observe aussi que les valeurs censurées précédant t_1 (la plus petite valeur distincte non censurée) ne sont pas prises en compte. Enfin, lorsque deux courbes de survie $\hat{S}_1(t)$ et $\hat{S}_2(t)$ correspondant à deux groupes distincts sont reportées sur un graphique et que $\hat{S}_2(t) > \hat{S}_1(t)$, alors les sujets du groupe 2 ont une durée de vie plus longue que ceux du groupe 1.

7.4 Le modèle des risques proportionnels de Cox

7.4.1 Position du problème

Soient $Y = T$ une variable "dépendante" de durée de vie et $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ un vecteur de covariables comme dans tout problème de régression. Aucune restriction n'est faite sur le nombre et la nature des variables

"indépendantes". En pratique, la matrice d'observations présente une caractéristique particulière à cause de l'indicateur de censure C :

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} (t_1, c_1) & x_{11} & \dots & x_{1p} \\ (t_2, c_2) & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ (t_n, c_n) & x_{n1} & \dots & x_{np} \end{pmatrix} \tag{7.4}$$

où $y_i = (t_i, c_i)$ avec $c_i = 0$ pour une durée de vie exacte ($T = t_i$) et $c_i = 1$ pour une donnée censurée ($T > t_i$) ($i = 1, \dots, n$).

Si on veut "régresser" la variable Y sur le vecteur $\underset{\sim}{X}$ afin de voir s'il existe une association entre les deux, on ne peut réellement appliquer la méthode de régression multiple ou de la régression logistique car celles-ci ne peuvent prendre en compte les données censurées. Il convient cependant de noter que s'il n'y avait pas de valeurs censurées, on pourrait faire une régression multiple de $Y^* = \log T$ sur $X_1, \dots, X_p$, la transformation logarithmique étant appliquée pour atténuer la dissymétrie à droite de la distribution de T et ainsi la normaliser (c'est-à-dire la rendre plus proche d'une loi gaussienne).

De nos jours, le problème de régression d'une durée de vie T sur un vecteur de covariables $\underset{\sim}{X}$ se résout par la méthode des risques proportionnels de Cox.

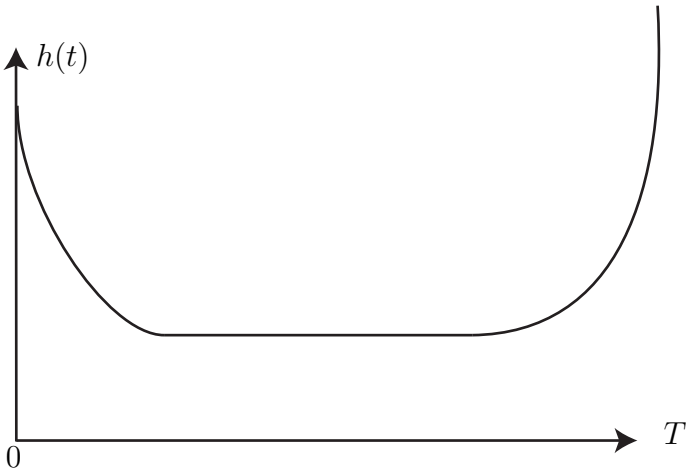
7.4.2 Fonction de risque

Soit $S(t) = P[T > t]$ la fonction de survie associée à la variable T. La *fonction de risque* ("hazard function"), notée $h(t)$, mesure le risque instantané de décès au temps t. Pour cela, il faut avoir vécu au moins jusqu'au temps $T = t$. Par définition,

$$h(t) = \frac{f(t)}{S(t)} \tag{7.5}$$

où $f(t)$ est la densité de probabilité de la variable T.

Chez l'être humain, la fonction $h(t)$ a la forme d'une baignoire (bathtub shape). Le risque de décéder est élevé à la naissance, il diminue ensuite pour atteindre un plateau constant à l'âge adulte, enfin il réaugmente rapidement dans la vieillesse pour tendre vers l'infini.



Puisque $f(t) = -\dfrac{dS(t)}{dt}$, l'équation (7.5) peut aussi s'écrire

$$h(t) = -\frac{d}{dt} \log S(t) \tag{7.6}$$

Si on veut exprimer la fonction de survie en termes de la fonction de risque, on montre facilement que

$$S(t) = e^{-\int_0^t h(u)du} \tag{7.7}$$

L'intégrale dans l'équation (7.7) est souvent appelée la fonction de risque cumulé $H(t) = \int_0^t h(u)du$. Dès lors,

$$S(t) = e^{-H(t)} \tag{7.8}$$

Ces relations sont importantes pour définir le modèle de Cox.

7.4.3 Les risques proportionnels

L'idée de David R. Cox (1972) fut d'introduire la notion de *risques proportionnels* ("proportional hazards, PH"). Considérons la variable T dans deux groupes de sujets et notons $h_1(t)$ et $h_2(t)$ les fonctions de risque des deux groupes. Supposons que ces risques soient proportionnels.

$$h_1(t) = \lambda h_2(t)$$

ou

$$\frac{h_1(t)}{h_2(t)} = \lambda \tag{7.9}$$

Si $\lambda = 1$, les fonctions de risque sont égales. Si $\lambda > 1$, les sujets du groupe 1 sont à plus haut risque que ceux du groupe 2, et l'inverse a lieu si $\lambda < 1$.

En utilisant les équations (7.7) et (7.9) on montre immédiatement que

$$S_1(t) = [S_2(t)]^{\lambda} \tag{7.10}$$

Dès lors, si $\lambda = 1$, les courbes de survie sont égales. Par contre, si $\lambda > 1, S_1(t) < S_2(t)$ et si $\lambda < 1, S_1(t) > S_2(t)$, corroborant ce qui a été dit plus haut.

L'examen de l'équation (7.10) montre que si on a une courbe de survie $S_0(t)$, en faisant varier λ, on peut définir une famille de courbes de survie (proportionnelles ou parallèles) au-dessus et en-dessous de $S_0(t)$ selon que λ est inférieur ou supérieur à 1. Ceci s'écrit

$$S(t|\lambda) = [S_0(t)]^{\lambda} \tag{7.11}$$

Comme λ ne peut être négatif, mais que le seuil $\lambda = 1$ est important, l'idée de Cox a été d'exprimer λ comme l'exponentielle d'une autre quantité, disons γ, soit $\lambda = e^{\gamma}$ où γ peut être négatif, nul ou positif. Dès lors, l'équation (7.11) s'écrit

$$S(t|\gamma) = [S_0(t)]^{e^{\gamma}} \qquad \gamma \in R \tag{7.12}$$

7.4.4 Régression de Cox

Le *modèle des risques proportionnels de Cox* permet alors d'exprimer la relation entre une variable de survie T et un vecteur de covariables $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$. En effet, si on pose $\underset{\sim}{\beta}^T \underset{\sim}{x} = \beta_1 x_1 + \ldots + \beta_p x_p$ comme on l'a toujours fait précédemment en régression, et qu'on se rappelle que cette somme pondérée peut prendre n'importe quelle valeur de $-\infty$ à $+\infty$, ce qui est précisément ce que fait γ dans l'expression (7.12), on peut écrire le modèle des risques proportionnels de Cox

$$S(t|\underset{\sim}{x}) = [S_0(t)]^{e^{\underset{\sim}{\beta}^T \underset{\sim}{x}}} \tag{7.13}$$

Ce modèle complexe montre comment varie la durée de vie des sujets en fonction des covariables envisagées.

On constate immédiatement que si $\underset{\sim}{\beta} = \underset{\sim}{0}, e^{\underset{\sim}{\beta}^T \underset{\sim}{x}} = 1$ et il n'y a pas d'association entre T et $\underset{\sim}{X}$. De même, on peut se poser la question de savoir si

$\beta_i = 0$, c'est-à-dire si la variable X_i est utile dans le modèle. On voit aussi que si l'index de risque $\underset{\sim}{\beta}^T \underset{\sim}{x}$ est négatif, la courbe de survie est plus favorable que la courbe de base $S_0(t)$. Par contre, si $\underset{\sim}{\beta}^T \underset{\sim}{x}$ est positif, la courbe de survie est plus défavorable. Donc, plus $\underset{\sim}{\beta}^T \underset{\sim}{x}$ est positif, plus la survie se détériore.

Un exemple particulièrement simple est le cas d'une seule variable binaire $X = 0, 1$. Dans ce cas, le modèle (7.13) s'écrit

$$S(t|x) = [S_0(t)]^{e^{\beta x}}$$

dont les deux seules courbes de survie s'écrivent $S_0(t)$ pour $x = 0$ et $[S_0(t)]^\gamma$ où $\gamma = e^\beta$ pour $x = 1$. On utilise souvent ce modèle pour comparer deux courbes de survie.

Notons enfin que si on se réfère à la fonction de risque, le modèle de Cox s'écrit

$$h(t|\underset{\sim}{x}) = h_0(t) \cdot e^{\underset{\sim}{\beta}^T \underset{\sim}{x}} \tag{7.14}$$

Le risque de base est multiplié par une fonction exponentielle qui incorpore les variables indépendantes $X_1, \ldots, X_p$.

7.5 Estimation des coefficients de régression

L'estimation des coefficients de régression $\underset{\sim}{\beta}^T = (\beta_1, \ldots, \beta_p)$ du modèle PH de Cox à partir de la matrice d'observations (7.4) n'est pas une sinécure et on doit à Cox l'ingéniosité d'avoir pu trouver la solution en introduisant une nouvelle forme de la vraisemblance, appelée vraisemblance "partielle" (partial likelihood). Celle-ci s'écrit

$$L(\underset{\sim}{\beta}) = \prod_{j=1}^{k^*} \left\{ \frac{e^{\underset{\sim}{\beta}^T \underset{\sim}{x}_j}}{\sum_{m \in R_j} e^{\underset{\sim}{\beta}^T \underset{\sim}{x}_m}} \right\} \tag{7.15}$$

où k^* est le nombre de temps de décès réels, R_j l'ensemble des sujets en vie juste avant t_j.

Le logarithme de la vraisemblance (7.15) s'écrit

$$l(\underset{\sim}{\beta}) = \sum_{j=1}^{k^*} \left\{ \underset{\sim}{\beta}^T \underset{\sim}{x}_j - \log \sum_{m \in R_j} e^{\underset{\sim}{\beta}^T \underset{\sim}{x}_m} \right\} \tag{7.16}$$

Pour trouver $\hat{\underset{\sim}{\beta}} = \underset{\sim}{b}$, il suffit de maximiser cette fonction. Ceci se fait par ordinateur et on obtient aussi les erreurs types des estimations, $SE(b_1), \ldots, SE(b_p)$, comme en régression logistique.

Connaissant le vecteur $\underset{\sim}{b}$, les équations (7.13) et (7.14) deviennent

$$\hat{S}(t|\underset{\sim}{x}) = \left[\hat{S}_0(t)\right]^{e^{\underset{\sim}{b}^T \underset{\sim}{x}}} \tag{7.17}$$

et

$$\hat{h}(t|\underset{\sim}{x}) = \hat{h}_0(t) \cdot e^{\underset{\sim}{b}^T \underset{\sim}{x}}$$

Pour estimer $\hat{S}_0(t)$, on se sert de l'estimation de la fonction de risque cumulé

$$\hat{H}_0(t) = \sum_{t_j \leq t} \left[\frac{d_j}{\sum_{m \in R_j} e^{\underset{\sim}{b}^T \underset{\sim}{x}_m}} \right] \tag{7.18}$$

et du fait que $\hat{S}_0(t) = e^{-\hat{H}_0(t)}$ pour toute valeur de t. Dès lors, pour un individu donné $\underset{\sim}{X} = \underset{\sim}{x}$, la probabilité de survie en $T = t$ est calculée par la formule

$$\begin{aligned} \hat{S}(t|\underset{\sim}{x}) &= \left[\hat{S}_0(t)\right]^{e^{\underset{\sim}{b}^T \underset{\sim}{x}}} \\ &= \left[e^{-\hat{H}_0(t)}\right]^{e^{\underset{\sim}{b}^T \underset{\sim}{x}}} \end{aligned} \tag{7.19}$$

où $\hat{H}_0(t)$ est donné en (7.18).

7.6 Tests sur le modèle

En général, on ne va pas jusqu'à calculer la probabilité de survie d'un individu donné. On se contente de tester des hypothèses sur le modèle comme on l'a fait pour la régression logistique.

Pour tester l'hypothèse $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$, on a recours au test du rapport de vraisemblance distribué comme un chi-carré à p degrés de liberté, comme on l'a fait précédemment. On peut aussi utiliser un test de Wald en calculant la forme quadratique (distance de Mahalanobis) $\chi^2_p = \underset{\sim}{b}^T \cdot \underset{\sim}{V}(b)^{-1} \cdot \underset{\sim}{b}$, où $\underset{\sim}{V}(b)$ est la matrice de variances-covariances asymptotique estimée du vecteur $\underset{\sim}{b}$. Il suffit de calculer ce critère distribué comme un chi-carré à p degrés de liberté et de le comparer au seuil critique $Q_{\chi^2}(1-\alpha; p)$.

7.6.1 Utilité des variables

Pour tester les hypothèses $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ (utilité de la variable X_i dans le modèle), on calcule le critère

$$Z = \frac{b_i}{SE(b_i)} \quad (i = 1, \ldots, p) \tag{7.20}$$

ou son carré Z^2 que l'on compare au seuil critique à 5% du chi-carré à 1 degré de liberté égal à 3.84. On rejette l'hypothèse nulle si Z^2 est supérieur à ce seuil.

7.6.2 Rapport de risque

Pour chaque variable X_i, on a estimé le coefficient de régression b_i et son erreur type $SE(b_i)$. Par définition, le *rapport de risque* ("hazard ratio") associé à la variable X_i s'écrit

$$\hat{h}_i = e^{b_i} \tag{7.21}$$

On peut calculer un intervalle de confiance pour le rapport de risque réel (inconnu) $h_i = e^{\beta_i}$ comme suit : on détermine d'abord l'intervalle de confiance pour β_i en calculant $b_i \pm 2\ SE(b_i)$. En prenant l'exponentielle de ces deux valeurs b_{i1} et b_{i2}, on obtient l'intervalle de confiance à 95% $h_{i1} \leq h_i \leq h_{i2}$. Cette façon de procéder permet de voir si h_i est différent de 1. Il suffit de voir si l'intervalle de confiance recouvre la valeur 1. Toutefois, ceci revient à voir si $\beta_i = 0$. Si $\hat{h}_i$ est significativement plus grand (plus petit) que 1, la variable X_i augmente (diminue) le risque de décès instantané (voir (7.17)). Cette interprétation est souvent utile lorsque toutes les variables $X_1, \ldots, X_p$ sont binaires. Elle permet de voir les variables qui augmentent le risque de décès et celles qui le diminuent.

7.7 Méthodes de sélection de variables

Comme en régression multiple et en régression logistique, on peut avoir recours à une méthode de sélection de variables ascendante, descendante ou "stepwise" pour ne retenir que les variables utiles dans le modèle. La sélection et le critère d'arrêt sont basés sur le maximum de la fonction de vraisemblance et le test du chi-carré. Dans la méthode ascendante, on s'arrête lorsque plus aucune des variables n'améliore significativement le modèle. Dans la méthode descendante, on s'arrête lorsque le retrait de toute variable non éliminée détériore de façon significative le modèle.

7.8 Exemple

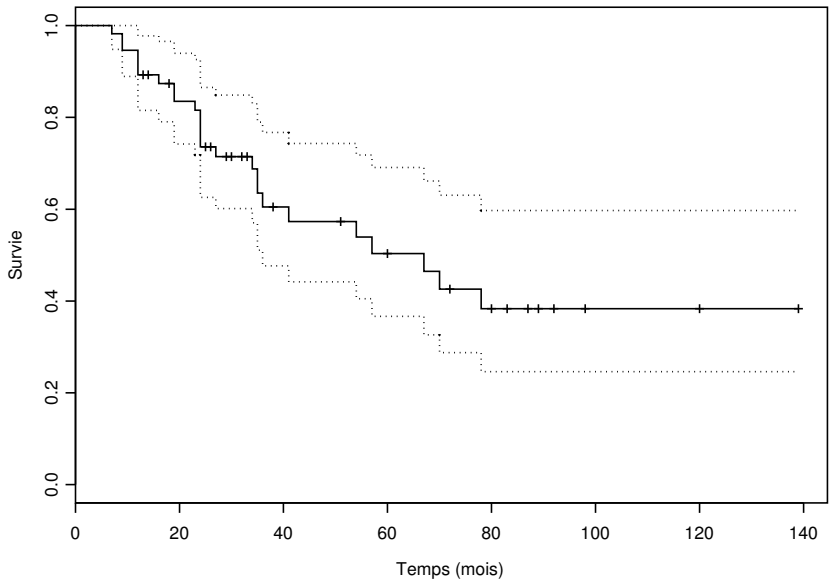
A titre d'illustration, considérons la durée de vie postopératoire (Y, mois) de 56 patients atteints d'un cancer rectal préalablement traités par radiothérapie. Les covariables envisagées sont : X_1=dose d'irradiation reçue (0 si <5000 rads ou 1 si $\geq$ 5000 rads), X_2=l'âge au moment de l'opération (années) et X_3 le sexe du patient (0=femme, 1=homme). Les données sont reprises à l'Annexe V.

7.8.1 Courbe de survie de Kaplan-Meier

On constate en triant les 56 durées de vie par ordre croissant que seules $k = 17$ d'entre elles correspondent à des durées non censurées distinctes. On établit d'abord la table de survie comme décrit précédemment.

Durée de vie t_i	En vie l_i	Décès d_i	Survie $\hat{S}(t_i)$
0	56	0	1.0
7	56	1	0.9821
9	53	2	0.9464
12	50	3	0.8929
16	46	1	0.8739
19	43	2	0.8350
23	42	1	0.8156
24	37	4	0.7360
27	33	1	0.7144
34	26	1	0.6879
35	24	2	0.6350
36	20	1	0.6048
41	18	1	0.5729
54	16	1	0.5392
57	14	1	0.5033
67	12	1	0.4646
70	11	1	0.4259
78	9	1	0.3833

La courbe de Kaplan-Meier avec sa région de confiance est représentée sur la figure ci-dessous.



7.8.2 Modèle PH de Cox

Parmi les 56 patients, 21 ont reçu une dose <5000 rads et 35 une dose supérieure ou égale à 5000 rads. Il y a 25 femmes et 31 hommes, d'âge moyen 62.1 ± 12 ans. La médiane est égale à 62 ans.

L'application de la régression de Cox de la durée de vie sur les trois covariables conduit aux résultats suivants :

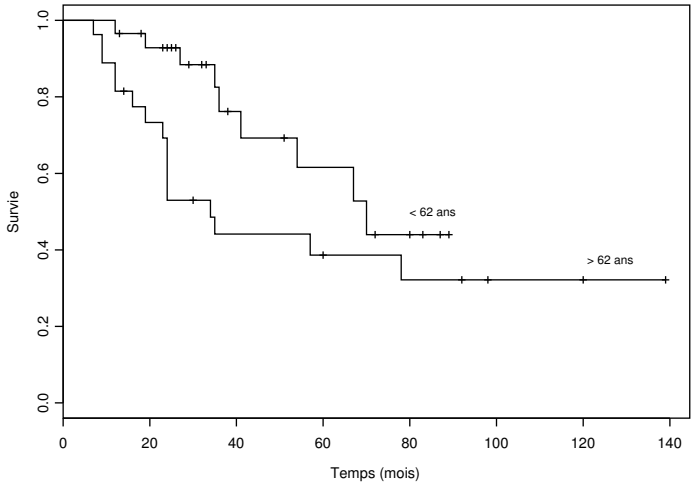
Variable	Estimation	Erreur type	Z^2	p-value	h
Dose	$b_1 = 0.08668$	0.4983	0.03	0.86	1.09
Age	$b_2 = 0.04786$	0.0201	5.68	0.017	1.05
Sexe	$b_3 = -0.8987$	0.4718	3.63	0.057	0.41

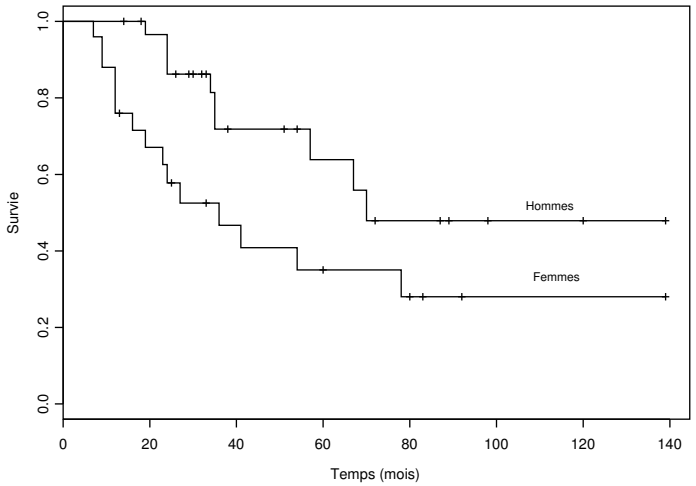
On constate que la dose d'irradiation n'a pas d'effet sur la durée de survie (p=0.86). Par contre, l'âge est un facteur significatif (p=0.017) dans la mesure où la survie est diminuée chez les patients plus âgés (coefficient positif). Le rapport de risque est égal à 1.05 par année supplémentaire au moment du diagnostic. Enfin, on observe un effet lié au sexe à la limite de la signification statistique (p=0.057). Le coefficient de régression étant négatif, le risque est diminué chez les hommes.

Le test du rapport de vraisemblance montre que la régression de Cox a un sens puisque $LR = 11.17$ à 3 degrés de liberté (p=0.011). De même, le test de Wald vaut 10.36 à 3 degrés de liberté (p=0.016).

L'application d'une méthode de sélection de variables ascendante pas à pas confirme que seuls l'âge (p=0.012) et le sexe (p=0.038) sont des facteurs de risque significatifs. Sur base des deux variables retenues, l'index de risque s'écrit $b^T x = 0.046 \times$ âge $- 0.86 \times$ sexe et le test du rapport de vraisemblance vaut $LR = 11.14$ à 2 degrés de liberté (p=0.0038). Le rapport de risque est égal à 1.05 pour l'âge et 0.42 pour le sexe ; ces valeurs sont proches de celles obtenues en incluant les 3 covariables dans le modèle.

Les figures ci-dessous illustrent l'effet de l'âge sur la durée de vie en comparant les sujets d'âge inférieur (n=29) et supérieur (n=27) à la médiane, ainsi que l'effet du sexe (femmes versus hommes).





Chapitre 8

Analyse discriminante

8.1 Introduction

L'analyse discriminante fait partie des méthodes les plus anciennes et les plus importantes de la statistique multivariée. En général, on considère deux ou plusieurs populations de sujets (ou d'objets) et un ensemble de variables. On se pose la question de savoir si les variables "discriminent", c'est-à-dire différencient les populations. Si c'est le cas, les variables sont discriminantes. En médecine, les populations sont des pathologies (des maladies) et les variables des tests cliniques, radiologiques ou biologiques. Ceux-ci permettent-ils de caractériser et de discerner les maladies ? On parle de diagnostic différentiel des maladies. En archéologie, des mesures effectuées sur des objets trouvés sont-elles capables de distinguer les époques auxquelles ces objets furent fabriqués ? En marketing, peut-on différencier différents types de clients sur base d'un questionnaire d'enquête ? Les exemples sont nombreux et peuvent se trouver dans tous les domaines de la vie.

De nos jours, l'analyse discriminante est utilisée comme méthode de classement de sujets parmi plusieurs populations. Comment classer un sujet d'origine inconnue parmi plusieurs populations sur base d'une observation multivariée avec un risque minimum de se tromper ?

On commencera par poser le problème de l'analyse discriminante dans le cas de deux groupes. On introduira la fonction discriminante de Fisher ainsi que les notions de probabilités a priori et a posteriori, de taux d'erreur et de classement correct. On montrera la relation entre fonction discriminante et régression logistique. On évoquera le problème de sélection de variables discriminantes. On envisagera ensuite le cas de plus de deux populations. Celui-ci peut être approché par la méthode d'analyse discriminante canonique, une approche similaire à l'analyse en composantes principales. Le problème peut

aussi être abordé par la construction de plusieurs fonctions discriminantes qui permettent de calculer les probabilités a posteriori. On illustrera le chapitre par différents exemples.

8.2 Discrimination entre deux groupes

Soient deux groupes (ou populations), notées H_1 et H_2, et un vecteur p-varié $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$. Est-il possible de discriminer entre les deux groupes sur base de $\underset{\sim}{X}$? Peut-on classer un sujet d'origine inconnue dans H_1 ou H_2 sur base de la seule observation multivariée $X = \underset{\sim}{x}$? Telles sont les questions que l'on se pose en analyse discriminante. Notons $\underset{\sim}{\mu}_1$ et $\underset{\sim}{\mu}_2$ les vecteurs moyens théoriques des deux populations et $\underset{\sim}{\sum}$ la matrice de variances-covariances supposée la même dans les deux populations (condition d'homoscédasticité).

Supposons que l'on dispose d'un échantillon d'effectif n_1 extrait du groupe H_1 et d'un échantillon d'effectif n_2 extrait du groupe H_2. On se trouve en présence de deux échantillons séparés ! L'observation du vecteur $\underset{\sim}{X}$ dans ces deux échantillons donnent lieu à deux matrices d'observations qu'on note $\underset{\sim}{X}_{n_1 \times p}$ et $\underset{\sim}{X}_{n_2 \times p}$ comme précédemment. Ces deux matrices d'observations définissent dans $\mathbb{R}^p$ deux nuages de points (voir Chapitre 2). Répondre aux questions ci-dessus revient à voir, en quelque sorte, si les deux nuages de points sont bien séparés ou se superposent de façon importante. Une autre approche consisterait à trouver un hyperplan qui les sépare de façon plus ou moins satisfaisante.

8.3 Test T^2 de Hotelling

Pour voir si le vecteur $\underset{\sim}{X}$ discrimine les groupes H_1 et H_2, il faut d'abord comparer les vecteurs moyens de $\underset{\sim}{X}$ dans les deux groupes. Si ceux-ci ne sont pas statistiquement différents, il n'y a pas lieu de poursuivre le problème car les nuages de points se superposent. Dans le cas inverse, on peut rechercher une fonction discriminante qui les sépare.

Pour tester l'hypothèse d'égalité des moyennes théoriques du vecteur $\underset{\sim}{X}$ dans les deux groupes H_1 et H_2, c'est-à-dire $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ versus $H_a : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$, on utilise le test T^2 de Hotelling. Il faut néanmoins supposer que le vecteur $\underset{\sim}{X}$ est constitué de variables continues normales.

A partir de la matrice d'observations $\underset{\sim}{X}_{n_1 \times p}$ du groupe H_1, on calcule (voir Chapitre 3) le vecteur moyen $\underset{\sim}{\bar{x}}_1$ et la matrice de dispersion $\underset{\sim}{S}_1$. On

procède de même à partir de la matrice d'observation $\underset{\sim}{X}_{n_2 \times p}$ du groupe H_2, soient $\underset{\sim}{\bar{x}}_2$ et $\underset{\sim}{S}_2$ le vecteur moyen et la matrice de dispersion.

On calcule alors la matrice de dispersion pondérée,

$$\underset{\sim}{S} = \frac{(n_1 - 1)\underset{\sim}{S}_1 + (n_2 - 1)\underset{\sim}{S}_2}{n_1 + n_2 - 2} \tag{8.1}$$

et ensuite le critère de Hotelling

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2) \tag{8.2}$$

où on reconnaît à droite la distance de Mahalanobis entre les deux moyennes $\underset{\sim}{\bar{x}}_1$ et $\underset{\sim}{\bar{x}}_2$.

Pour tester l'égalité des moyennes $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$, on calcule ensuite la quantité

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 \tag{8.3}$$

distribuée comme une loi F de Snedecor à p et $n_1 + n_2 - p - 1$ degrés de liberté.

On rejette H_0 si $F \geq Q_F(1 - \alpha; p, n_1 + n_2 - p - 1)$ le quantile $1 - \alpha$ du F de Snedecor à p et $n_1 + n_2 - p - 1$ degrés de liberté, sinon on ne rejette pas H_0. Dans le premier cas, on peut dire que le vecteur $\underset{\sim}{X}$ discrimine H_1 et H_2 et on peut rechercher une règle pour classer les sujets dans l'un ou l'autre groupe.

8.4 Fonction linéaire discriminante de Fisher

Lorsque la séparation entre les groupes H_1 et H_2 est suffisante, on peut rechercher dans l'espace à p dimensions la direction qui sépare au mieux les deux nuages de points.

On a vu que pour passer de l'espace à p-dimensions à l'espace à 1 dimension (droite réelle), il suffisait de prendre une combinaison linéaire $\underset{\sim}{\beta}^T \underset{\sim}{X} = \beta_1 X_1 + \ldots + \beta_p X_p$ du vecteur $\underset{\sim}{X}$. Si on calcule $\underset{\sim}{\beta}^T \underset{\sim}{X}$ pour toutes les observations de la matrice d'observations $\underset{\sim}{X}_{n_1 \times p}$ du groupe 1 et celles $\underset{\sim}{X}_{n_2 \times p}$ du groupe 2, on obtient deux échantillons univariés dont les moyennes valent respectivement $\underset{\sim}{\beta}^T \underset{\sim}{\bar{x}}_1$ et $\underset{\sim}{\beta}^T \underset{\sim}{\bar{x}}_2$, tandis que la variance pondérée vaut $\underset{\sim}{\beta}^T \underset{\sim}{S} \underset{\sim}{\beta}$ où $\underset{\sim}{S}$ est définie en (8.1). On va rechercher la direction (c'est-à-dire le vecteur $\underset{\sim}{\beta}$) qui maximise la distance entre les deux moyennes, soit $[\underset{\sim}{\beta}^T \underset{\sim}{\bar{x}}_1 - \underset{\sim}{\beta}^T \underset{\sim}{\bar{x}}_2]^2 =$

$[\underset{\sim}{\beta}^T(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)]^2$ en tenant compte de la variance à l'intérieur de chaque groupe $\underset{\sim}{\beta}^T \underset{\sim}{S} \underset{\sim}{\beta}$. En d'autres termes, on cherche le vecteur $\underset{\sim}{\beta}$ qui maximise le rapport

$$\lambda = \frac{[\underset{\sim}{\beta}^T(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)]^2}{\underset{\sim}{\beta}^T \underset{\sim}{S} \underset{\sim}{\beta}} \tag{8.4}$$

Il est facile de montrer que cette direction est égale à $\underset{\sim}{\hat{\beta}} = \underset{\sim}{b} = \underset{\sim}{S}^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)$ où tout est connu.

Dès lors, la combinaison linéaire

$$\begin{aligned} L(\underset{\sim}{x}) &= \underset{\sim}{b}^T \underset{\sim}{x} \\ &= \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)^T \underset{\sim}{S}^{-1} \underset{\sim}{x} \end{aligned} \tag{8.5}$$

est appelée "fonction linéaire discriminante de Fisher". Elle fournit la meilleure séparation entre les deux nuages de points de l'espace à p-dimensions.

8.5 Règle de classement

Pour classer un nouvel individu $\underset{\sim}{x}$ dans le groupe H_1 ou le groupe H_2 sur base de la fonction linéaire discriminante de Fisher $L(\underset{\sim}{x})$, il faut un seuil de décision (cut-off point). On pourrait prendre le milieu des valeurs moyennes de $L(\underset{\sim}{x})$ dans H_1 et H_2, c'est-à-dire le seuil

$$\begin{aligned} c &= \frac{L(\underset{\sim}{\bar{x}}_1) + L(\underset{\sim}{\bar{x}}_2)}{2} \\ &= \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)^T \underset{\sim}{S}^{-1} \left(\frac{\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2}{2}\right) \end{aligned} \tag{8.6}$$

Dès lors, on classe un sujet $\underset{\sim}{x}$ dans H_1 si $L(\underset{\sim}{x}) \geq c$ et dans H_2 si $L(\underset{\sim}{x}) < c$. Si on retire la constante c de $L(x)$, la fonction discriminante de Fisher s'écrit

$$L^*(\underset{\sim}{x}) = \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right)\right] \tag{8.7}$$

et la règle s'énonce plus simplement :

$$\begin{aligned} &\text{"Classer } \underset{\sim}{x} \text{ dans } H_1 \text{ si } L^*(\underset{\sim}{x}) \geq 0\text{"} \\ &\text{"Classer } \underset{\sim}{x} \text{ dans } H_2 \text{ si } L^*(\underset{\sim}{x}) < 0\text{"} \end{aligned} \tag{8.8}$$

L'équation $L^*(\underset{\sim}{x}) = 0$ définit dans l'espace des observations $\mathbb{R}^P$ un hyperplan de séparation des deux nuages de points.

8.6 Taux d'erreur

Afin d'évaluer la qualité de la règle de classement, donc de la fonction linéaire discriminante de Fisher, on calcule pour chaque observation $\underset{\sim}{x}_i$ du groupe H_1 et du groupe H_2 la valeur $L^*(\underset{\sim}{x}_i)$ définie en (8.7). Une observation du groupe H_1 devrait avoir $L^*(\underset{\sim}{x}_i) \geq 0$ et une observation du groupe H_2 $L^*(\underset{\sim}{x}_i) < 0$. Si tel n'est pas le cas, alors on peut considérer l'observation comme une erreur de classement. Deux erreurs sont possibles : l'observation $\underset{\sim}{x}_i$ provient du groupe H_1 mais $L^*(\underset{\sim}{x}_i) < 0$ et est donc classée dans le groupe H_2 (1e erreur) ou l'observation $\underset{\sim}{x}_i$ provient du groupe H_2 mais $L^*(\underset{\sim}{x}_i) \geq 0$ et est donc classée dans le groupe H_1 (2e erreur). Soient r_1 le nombre d'erreurs pour le groupe H_1 et r_2 le nombre d'erreurs pour le groupe H_2. Dans ces conditions, le taux d'erreur par resubstitution de la règle de classement (8.8) vaut

$$\varepsilon_R = \frac{r_1 + r_2}{n_1 + n_2} \tag{8.9}$$

8.7 Probabilités a priori et a posteriori

Si on considère le problème d'analyse discriminante comme un problème de classement, il est nécessaire d'introduire les notions de probabilités a priori et a posteriori.

Par définition, les probabilités a priori $\pi_1 = P(H_1)$ et $\pi_2 = P(H_2)$ sont les probabilités d'appartenir aux groupes H_1 et H_2 avant l'observation du vecteur $\underset{\sim}{X}$. Elles représentent en quelque sorte les proportions des deux groupes dans le mélange. On a $\pi_1 + \pi_2 = 1$. Par exemple si $\pi_1 = 0.10$ et $\pi_2 = 0.90$, en tirant un sujet au hasard, il n'a que 10% de chances d'appartenir à H_1. C'est un élément qui doit être pris en compte dans le classement du sujet après avoir observé $\underset{\sim}{X} = \underset{\sim}{x}$.

Les probabilités a posteriori sont les probabilités d'appartenir aux groupes H_1 et H_2 après avoir observé le vecteur $\underset{\sim}{X}$. On les note

$$P_1(\underset{\sim}{x}) = P[H_1|\underset{\sim}{x}]$$

et (8.10)

$$P_2(\underset{\sim}{x}) = P[H_2|\underset{\sim}{x}]$$

et on a toujours $P_1(\underset{\sim}{x}) + P_2(\underset{\sim}{x}) = 1$.

Sous certaines hypothèses de normalité, il est aisé de montrer que

$$P_1(\underset{\sim}{x}) = \frac{e^{\log \frac{\pi_1}{\pi_2} + L^*(\underset{\sim}{x})}}{1 + e^{\log \frac{\pi_1}{\pi_2} + L^*(\underset{\sim}{x})}}$$

et (8.11)

$$P_2(\underset{\sim}{x}) = \frac{1}{1 + e^{\log \frac{\pi_1}{\pi_2} + L^*(\underset{\sim}{x})}}$$

où $L^*(\underset{\sim}{x})$ est la fonction discriminante définie en (8.7). On reconnaît la forme logistique des probabilités a posteriori.

Notons que la règle de classement (8.8) peut alors s'écrire

"Classer $\underset{\sim}{x}$ dans H_1 si $P_1(\underset{\sim}{x}) \geq P_2(\underset{\sim}{x})$
et dans H_2 sinon." (8.12)

Donc on classe le sujet dans le groupe pour lequel il a la plus grande probabilité a posteriori, ce qui semble logique. On voit que les probabilités a priori interviennent dans les formules (8.11).

En termes de fonction linéaire discriminante, la règle de classement (8.8) ou (8.12) peut aussi s'écrire

"Classer $\underset{\sim}{x}$ dans H_1 si $L^*(\underset{\sim}{x}) \geq \log \frac{\pi_2}{\pi_1}$
et dans H_2 sinon." (8.13)

Lorsque $\pi_1 = \pi_2$, on retrouve la règle de classement (8.8). On voit aussi que si $\pi_2 > \pi_1$ (a priori plus de chances d'appartenir à H_2), $\log \frac{\pi_2}{\pi_1} > 0$ et il faut que $L^*(\underset{\sim}{x})$ excède un seuil positif pour que $\underset{\sim}{x}$ soit classé dans H_1 (on est donc plus sévère qu'en utilisant le seuil nul). L'inverse se produit si $\pi_2 < \pi_1$.

8.8 Autres considérations

Comme dans tous les autres problèmes de statistique multivariée, il existe des techniques (ascendantes ou descendantes) qui permettent de sélectionner les meilleures variables discriminantes sans perte significative d'information. Elles sont basées sur un test F de Snedecor à 1 et $n-k-1$ degrés de liberté.

On a pu montrer que la fonction linéaire discriminante de Fisher pouvait être obtenue, à un facteur constant près, en calculant la régression multiple d'une variable artificielle Y égale à 0 pour H_1 et à 1 pour H_2. Ceci est intéressant car il suffit d'introduire les données dans un programme de régression multiple.

Enfin, on peut noter qu'il y a d'autres méthodes pour calculer le taux d'erreurs (ε) en plus de la méthode de resubstitution. Notons d'abord le cas $\underset{\sim}{X}$ multinormal pour lequel le taux d'erreur (par insertion) vaut

$$\varepsilon_I = F_G\left(-\frac{D}{2}\right) \tag{8.14}$$

où $F_G(t)$ est la fonction de répartition gaussienne, soit la fonction $P(Z \leq t)$ et D la racine carrée de la distance de Mahalanobis $(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)$.

L'autre approche est celle d'extraction-réinsertion (dite du Jackknife) qui consiste à retirer une observation de l'échantillon des $n = n_1 + n_2$ observations, de calculer la fonction discriminante de Fisher sur base des $n - 1$ observations restantes et de reclasser l'observation retirée. On répète ce processus pour toutes les $n_1 + n_2$ observations. Si on note r'_1 et r'_2 le nombre d'observations mal classées, le taux d'erreur s'écrit

$$\varepsilon_J = \frac{r'_1 + r'_2}{n_1 + n_2} \tag{8.15}$$

En général, ce taux d'erreur est plus réaliste car $\varepsilon_J \geq \varepsilon_R$.

8.9 Application

Afin d'illuster la discrimination entre 2 groupes et la fonction linéaire discriminante de Fisher, nous avons appliqué la méthode aux iris versicolor (groupe H_1) et virginica (groupe H_2) de Fisher sur base des 4 variables traditionelles : longueur des sépales (X_1), largeur des sépales (X_2), longueur des pétales (X_3) et largeur des pétales (X_4). Comme il y a 50 fleurs dans chaque groupe, $n = n_1 + n_2 = 100$. Le tableau ci-dessous donne la moyenne $\pm$ SD de chaque variable dans les deux groupes.

Variable	Iris versicolor $n_1 = 50$	Iris virginica $n_2 = 50$	$(\bar{x}_1 + \bar{x}_2)/2$
Longueur sépale	59.4 ± 5.16	65.9 ± 6.36	62.7
Largeur sépale	27.7 ± 3.14	29.7 ± 3.23	28.7
Longueur pétale	42.6 ± 4.70	55.5 ± 5.52	49.1
Largeur pétale	13.3 ± 1.98	20.3 ± 2.75	16.8

Les matrices de variances-covariances dans chaque groupe $\underset{\sim}{S}_1$ et $\underset{\sim}{S}_2$ et la matrice de variances-covariances pondérée s'écrivent successivement

$$\underset{\sim}{S}_1 = \begin{pmatrix} 26.64 & & & \\ 8.518 & 9.847 & & \\ 18.29 & 8.265 & 22.08 & \\ 5.578 & 4.120 & 7.310 & 3.911 \end{pmatrix}$$

$$\underset{\sim}{S}_2 = \begin{pmatrix} 40.43 & & & \\ 9.376 & 10.40 & & \\ 30.33 & 7.138 & 30.46 & \\ 4.909 & 4.763 & 4.882 & 7.543 \end{pmatrix}$$

$$\underset{\sim}{S} = \begin{pmatrix} 33.54 & & & \\ 8.947 & 10.12 & & \\ 24.31 & 7.702 & 26.27 & \\ 5.244 & 4.442 & 6.096 & 5.727 \end{pmatrix}$$

Après inversion de la matrice $\underset{\sim}{S}$, la distance de Mahalanobis entre les 2 groupes s'écrit

$$\begin{aligned} D^2 &= (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1} (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2) \\ &= 14.219 \end{aligned}$$

Le critère de Hotelling a pour valeur

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 = \frac{2500}{100} \times 14.219 = 355.48$$

ce qui conduit à une valeur F de Snedecor égale à

$$\begin{aligned} F &= \frac{50 + 50 - 4 - 1}{4 \times (100 - 2)} \times 355.48 \\ &= 86.15 \end{aligned}$$

Les degrés de liberté valent respectivement $\nu_1 = p = 4$ et $\nu_1 = n_1 + n_2 - p - 1 = 95$. Comme le quantile à 95% du F de Snedecor vaut

$$Q_F(0.95; 4, 95) = 2.47$$

on rejette l'hypothèse $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$. Dès lors, le vecteur $\underset{\sim}{X}^T = (X_1, X_2, X_3, X_4)$ des sépales et des pétales discrimine les deux espèces d'iris versicolor et virginica. On peut noter que la probabilité de dépassement associée à la valeur observée du F de Snedecor est $p < 0.0001$.

La fonction linéaire discriminante de Fisher s'écrit

$$\begin{aligned} L(\underset{\sim}{x}) &= (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1} \underset{\sim}{x} \\ &= 0.3556X_1 + 0.5579X_2 - 0.6970X_3 - 1.239X_4 \end{aligned}$$

Puisque $c = -16.66$, la fonction linéaire discriminante de Fisher corrigée devient

$$L^*(\underset{\sim}{x}) = 16.66 + 0.3556X_1 + 0.5579X_2 - 0.6970X_3 - 1.239X_4$$

Une fleur est classée dans le groupe des iris versicolor si $L^*(\underset{\sim}{x}) \geq 0$ et dans le groupe des iris virginica si $L^*(\underset{\sim}{x}) < 0$. A titre d'exemple, si on remplace le vecteur $\underset{\sim}{X}$ par les valeurs moyennes reprises dans le tableau ci-dessus pour le groupe des iris versicolor, on trouve effectivement une valeur positive. En effet,

$$\begin{aligned} L^*(\underset{\sim}{x}) &= 16.66 + 0.3556 \times 59.4 + 0.5579 \times 27.7 - 0.6970 \times 42.6 - 1.239 \times 13.3 \\ &= 7.11 \end{aligned}$$

Cette valeur n'est autre que la moitié de la distance de Mahalanobis. Par contre, si on y substitue les valeurs moyennes des iris virginica, on trouve $L^*(\underset{\sim}{x}) = -7.11$, soit la même valeur mais de signe opposé.

Lorsqu'on recalcule la fonction discriminante $L^*(\underset{\sim}{x})$ pour toutes les fleurs du groupe 1, on constate qu'une seule est mal classée ($r_1 = 1$). Par contre, pour les fleurs du groupe 2, le calcul montre que 2 sont mal classées ($r_2 = 2$). Dès lors, le taux d'erreur par resubstitution vaut

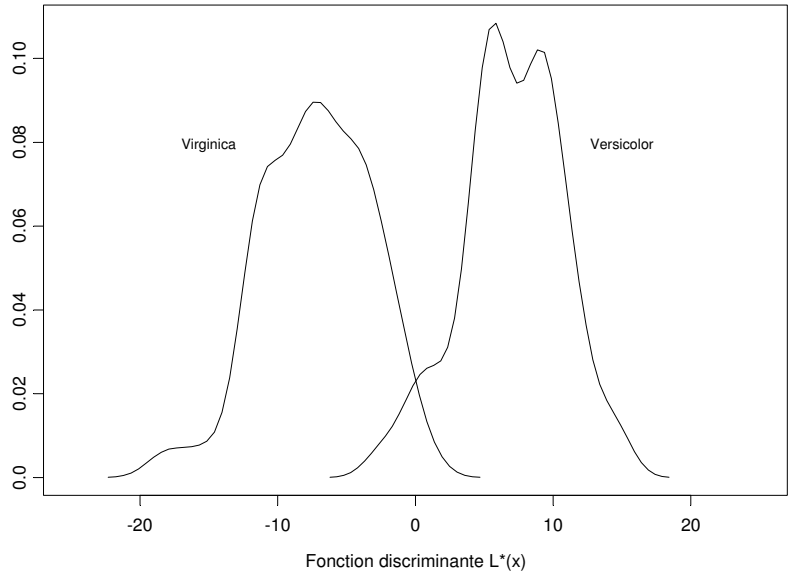
$$\varepsilon_R = \frac{1+2}{50+50} = 0.03$$

Le taux d'erreur par insertion vaut, puisque $D/2 = 1.8854$,

$$\varepsilon_I = F_G(-1.8854) = 0.0297,$$

soit une valeur très proche. Le taux d'erreur par extraction-réinsertion (jackknife) est inchangé puisque $r' = 1$ et $r'_2 = 2$.

A chaque fois, on a supposé que les probabilités a priori des deux groupes étaient égales $\pi_1 = \pi_2 = 0.50$. La figure ci-dessous, qui représente la densité estimée de $L^*(\underset{\sim}{x})$ dans chaque groupe, confirme que le degré de recouvrement des iris versicolor et virginica est faible.



8.10 Discrimination logistique

A la section 8.2.7, on a évoqué la relation entre l'analyse discriminante à 2 groupes et la régression multiple. De nos jours, il est souvent préféré de résoudre le problème d'analyse discriminante par la méthode de régression logistique. Il faut toutefois bien comprendre qu'en analyse discriminante la variable Y est fixée et le vecteur $\underset{\sim}{X}$ est observé alors que c'est l'inverse en régression. Anderson a montré qu'on pouvait néanmoins faire une régression logistique comme si Y avait été observée mais qu'il fallait apporter une correction au terme indépendant de la fonction discriminante, c'est-à-dire au seuil de décision pour tenir compte des probabilités a priori réelles. En effet, en faisant une régression logistique, tout se passe comme si $\pi_1 = n_1/n$ et $\pi_2 = n_2/n$, alors qu'en réalité ce n'est pas le cas. L'avantage de la régression logistique est qu'elle s'applique en toute généralité et que les probabilités a posteriori ont par définition la forme logistique.

Si on effectue une régression logistique de la variable binaire Y(0 pour H_1

et 1 pour H_2) sur le vecteur $\underset{\sim}{X}$, on obtient la fonction linéaire

$$\Lambda(\underset{\sim}{x}) = b_0 + b_1 x_1 + \ldots + b_p x_p \tag{8.16}$$

qui n'est pas exactement la même que la fonction linéaire discriminante (corrigée) de Fisher obtenue par régression multiple.

Si π_1 et π_2 sont les probabilités a priori réelles, on applique une correction au terme b_0 qui devient

$$b_0^* = b_0 + \log \frac{n_2 \pi_1}{n_1 \pi_2} \tag{8.17}$$

Dans ces conditions, la fonction logistique discriminante

$$\Lambda^*(\underset{\sim}{x}) = b_0^* + \underset{\sim}{b}^T \underset{\sim}{x} \tag{8.18}$$

conduit à la règle de classement suivante :

$$\begin{array}{l} \text{“Classer } \underset{\sim}{x} \text{ dans } H_1 \text{ si } \Lambda^*(\underset{\sim}{x}) \geq 0 \\ \text{et dans } H_2 \text{ sinon.”} \end{array} \tag{8.19}$$

Par ailleurs, les probabilités a posteriori s'écrivent

$$\begin{array}{l} P(H_1|\underset{\sim}{x}) = \dfrac{e^{\Lambda^*(\underset{\sim}{x})}}{1 + e^{\Lambda^*(\underset{\sim}{x})}} \\ \text{et} \\ P(H_2|\underset{\sim}{x}) = \dfrac{1}{1 + e^{\Lambda^*(\underset{\sim}{x})}} \end{array} \tag{8.20}$$

Le taux d'erreur par resubstitution se calcule aisément mais celui par Jackknife est plus fastidieux. Il n'y a aucune difficulté à appliquer les méthodes de sélection de variables décrites au Chapitre 6.

L'application de la méthode de régression logistique aux données des iris versicolor et virginica conduit aux résultats suivants :

$$\Lambda(\underset{\sim}{x}) = 42.64 + 0.2465 X_1 + 0.6681 X_2 - 0.9429 X_3 - 1.829 X_4$$

Puisque $\pi_1 = \pi_2$ et $n_1 = n_2$, le terme de correction s'annule et $b_0^* = b_0$. Dès lors $\Lambda^*(\underset{\sim}{x}) = \Lambda(\underset{\sim}{x})$. Le test du rapport de vraisemblance vaut $LR = 126.7$ à 4 degrés de liberté ; on observe donc une différence hautement significative entre les deux groupes ($p<0.0001$). Notons que la fonction discriminante logistique obtenue ici n'est pas tout à fait la même que celle obtenue en appliquant l'approche classique de Fisher. Le rapport entre les coefficients de ces deux équations varie entre 0.68 et 2.56 (valeur moyenne 1.46). Il est bien

connu qu'en théorie, la relation entre la fonction logistique et la fonction de répartition gaussienne fait intervenir un facteur égal à $\sqrt{8/\pi} = 1.60$. La valeur moyenne trouvée (1.46) en est fort proche.

Lorsque, $\Lambda(\underset{\sim}{x}) \geq 0$, l'observation est classée dans le groupe des iris versicolor et dans le groupe des iris virginica sinon.

8.11 Discrimination entre plusieurs groupes

8.11.1 Position du problème

Considérons à présent le problème de discrimination (et de classement) entre g groupes $H_1, \ldots, H_g$ sur base du vecteur $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$.

On peut aisément montrer que pour résoudre le problème, on a besoin de $g-1$ fonctions discriminantes, notées $L_1(\underset{\sim}{x}), \ldots, L_{g-1}(\underset{\sim}{x})$. On constate que si $g = 2$ (cas précédent), il faut une seule $(g - 1 = 1)$ fonction discriminante.

Supposons que l'on dispose d'une matrice d'observations de chaque groupe. Soient $\underset{\sim}{X}_{n_1 \times p}, \ldots, \underset{\sim}{X}_{n_g \times p}$, ces matrices où les effectifs $n_1, \ldots, n_g$ ont été fixés et $n = n_1 + \ldots + n_g$ est le nombre total d'observations. Pour chaque matrice d'observations, on peut calculer le vecteur moyen $\underset{\sim}{\bar{x}}_i$ et la matrice de dispersion $\underset{\sim}{S}_i$ $(i = 1, \ldots, g)$. Comme précédemment, on calcule la matrice pondérée

$$\underset{\sim}{S} = \frac{(n_1 - 1)\underset{\sim}{S}_1 + \ldots, +(n_g - 1)\underset{\sim}{S}_g}{n - g} \tag{8.21}$$

8.11.2 Analyse discriminante canonique

Dans l'espace à p dimensions, les g matrices d'observations forment g nuages de points qui se recouvrent dans certaines proportions. L'analyse discriminante canonique consiste à obtenir une représentation graphique aussi fidèle que possible des nuages de points de l'espace p-dimensionnel sur un plan à 2 dimensions (ou éventuellement dans un espace à 3 dimensions). Les axes sont choisis (comme pour la fonction linéaire discriminante de Fisher) de manière à maximiser les différences entres les moyennes des groupes (variabilité inter-groupes) par rapport à la variabilité à l'intérieur des groupes (variabilité intra-groupes).

La matrice des sommes de carrés et produits croisés inter-groupes a pour

expression

$$\underset{\sim}{H} = \sum_{j=1}^{g} \left(\underset{\sim}{\bar{x}}_j - \underset{\sim}{\bar{x}}\right)\left(\underset{\sim}{\bar{x}}_j - \underset{\sim}{\bar{x}}\right)^T \tag{8.22}$$

où $\underset{\sim}{\bar{x}} = (\underset{\sim}{\bar{x}}_1 + \ldots + \underset{\sim}{\bar{x}}_g)/g$, et la matrice des sommes de carrés et produits croisés intra-groupes vaut $\underset{\sim}{E} = (n-g)\underset{\sim}{S}$.

Pour résoudre le problème, il suffit de calculer les valeurs propres et vecteurs propres de la matrice $\underset{\sim}{E}^{-1}\underset{\sim}{H}$, donc de rechercher les racines de l'équation

$$\left|\underset{\sim}{E}^{-1}\underset{\sim}{H} - \lambda\, \underset{\sim}{I}\right| = 0 \tag{8.23}$$

Le rang de la matrice $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ est égal au rang de $\underset{\sim}{H}$ car $\underset{\sim}{E}$ est de rang maximum. Vu la définition (8.22), le rang de $\underset{\sim}{H}$ est inférieur ou égal à $\min(g-1, p)$ et vaut donc au plus $g-1$ pour autant que $p \geq g$. Soient $\lambda_1 > \ldots > \lambda_{g-1} \geq 0$, les $g-1$ valeurs propres par ordre décroissant.

Le premier axe (ou variable) canonique $Z_1 = \underset{\sim}{a}_1^T \underset{\sim}{X}$ correspond à la plus grande valeur propre λ_1 de la matrice $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ et $\underset{\sim}{a}_1$ est le vecteur propre correspondant normé à l'unité. Souvent, on corrige la première variable canonique en retirant la moyenne générale $\underset{\sim}{\bar{x}}$, soit $Z_1 = \underset{\sim}{a}_1^T(\underset{\sim}{x} - \underset{\sim}{\bar{x}})$.

Le deuxième axe (variable) canonique $Z_2 = \underset{\sim}{a}_2^T \underset{\sim}{X}$ correspond à la deuxième plus grande valeur propre de $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ et $\underset{\sim}{a}_2$ le vecteur propre orthonormé correspondant.

On procède ainsi de suite avec les autres variables canoniques $Z_3, \ldots, Z_{g-1}$ mais en général, on se limite aux deux premières.

La qualité de la discrimination s'apprécie en calculant le rapport

$$Q = \frac{\lambda_1 + \lambda_2}{\operatorname{tr}\left(\underset{\sim}{E}^{-1}\underset{\sim}{H}\right)} \tag{8.24}$$

comme en analyse en composantes principales.

Les deux variables canoniques $Z_1 = \underset{\sim}{a}_1^T(\underset{\sim}{x} - \underset{\sim}{\bar{x}})$ et $Z_2 = \underset{\sim}{a}_2^T(\underset{\sim}{x} - \underset{\sim}{\bar{x}})$ permettent de reporter les points de tous les groupes sur un plan et en particulier leurs moyennes. On peut ainsi localiser les groupes les uns par rapport aux autres.

Si $\lambda_3 = 0$ alors toutes les autres valeurs propres restantes $\lambda_4, \ldots, \lambda_{g-1}$ sont aussi nulles et on obtient une représentation exacte des nuages sur le plan. C'est notamment le cas pour $g = 3$. Il y a toujours un plan qui passe par 3 points !

Dans le cas où $g = 2$, la première et seule variable canonique est la fonction linéaire discriminante de Fisher.

En analyse canonique discriminante, on peut aussi faire de la sélection de variables pour rechercher et ne retenir que celles qui sont réellement discriminantes.

A titre d'illustration, nous avons appliqué l'analyse discriminante canonique aux trois groupes d'iris de Fisher (H_1=setosa, H_2=versicolor, H_3=virginica). On dispose au total de 150 données et de 4 variables comme précedemment. Puisque $g = 3$, on aura une représentation exacte des données dans le plan canonique.

Le tableau ci-dessous reprend les moyennes ± SD des 4 variables dans les trois groupes ainsi que pour l'ensemble des données

Variable	Iris setosa ($n_1 = 50$)	Iris versicolor ($n_2 = 50$)	Iris virginica ($n_3 = 50$)	Global ($n = 150$)
Longueur sépale	50.1 ± 3.52	59.4 ± 5.16	65.9 ± 6.36	58.4 ± 8.28
Largeur sépale	34.3 ± 3.79	27.7 ± 3.14	29.7 ± 3.23	30.6 ± 4.36
Longueur pétale	14.6 ± 1.74	42.6 ± 4.70	55.5 ± 5.52	37.6 ± 17.7
Largeur pétale	2.46 ± 1.05	13.3 ± 1.98	20.3 ± 2.75	12.0 ± 7.62

La matrice de variances-covariances pondérée s'écrit

$$\underset{\sim}{S} = \begin{pmatrix} 26.50 & & & \\ 9.272 & 11.54 & & \\ 16.75 & 5.524 & 18.52 & \\ 3.84 & 3.271 & 4.267 & 4.188 \end{pmatrix}$$

Une analyse de la variance multivariée montre que l'hypothèse d'égalité des vecteurs moyens des trois groupes $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_3$ est rejetée ($p<0.0001$). Par conséquent, le vecteur $\underset{\sim}{X}$ discrimine les trois groupes de fleurs.

La matrice des sommes de carrés et produits croisés intra-groupes $\underset{\sim}{E} = (n - g)\underset{\sim}{S}$ s'écrit

$$\underset{\sim}{E} = \begin{pmatrix} 3896 & & & \\ 1363 & 1696 & & \\ 2462 & 812.1 & 2722 & \\ 564.5 & 480.8 & 627.2 & 615.7 \end{pmatrix}$$

et la matrice des sommes de carrés et produits croisés inter-groupes a pour expression

$$\underset{\sim}{H} = \begin{pmatrix} 6321 & & & \\ -1995 & 1135 & & \\ 16575 & -5724 & 43710 & \\ 7128 & -2293 & 18677 & 8041 \end{pmatrix}.$$

Dès lors la matrice $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ a pour valeur

$$\underset{\sim}{E}^{-1}\underset{\sim}{H} = \begin{pmatrix} -3.052 & -5.565 & 8.075 & 10.493 \\ 1.079 & 2.180 & -2.942 & -3.418 \\ -8.096 & -14.975 & 21.505 & 27.538 \\ -3.452 & -6.312 & 9.140 & 11.840 \end{pmatrix}.$$

On constate que cette matrice n'est pas symétrique. Par ailleurs, son rang est inférieur ou égal à $min(g-1, p) = min(2, 4) = 2$.

Les deux valeurs propres non nulles de $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ valent respectivement $\lambda_1 = 32.19$ et $\lambda_2 = 0.2854$. Sachant que la trace de la matrice vaut 32.477, la première valeur propre représente à elle seule 99.1% de la discrimination entre les trois groupes.

Le premier et le second axes canoniques s'écrivent respectivement

$$Z_1 = -0.083X_1' - 0.153X_2' + 0.220X_3' + 0.281X_4'$$

et

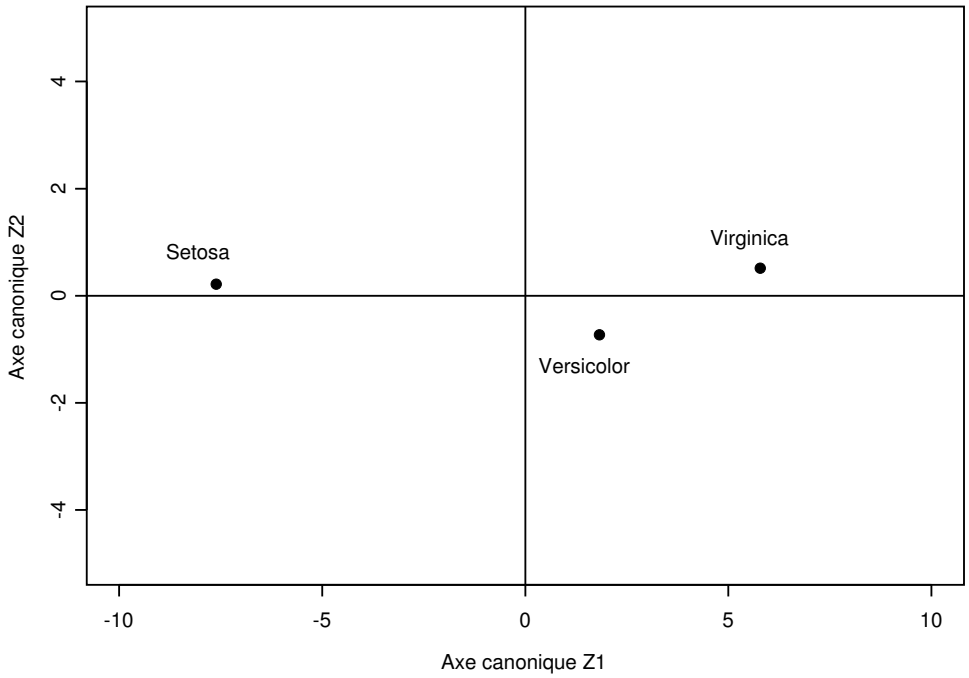
$$Z_2 = 0.002X_1' + 0.216X_2' - 0.093X_3' + 0.284X_4'$$

où les variables ont été préalablement centrées par rapport à la moyenne générale, c'est-à-dire $X_1' = (X_1 - 58.4), X_2' = (X_2 - 30.6), X_3' = (X_3 - 37.6), X_4' = (X_4 - 12.0)$.

Les vecteurs des moyennes des trois groupes projetés dans l'espace canonique à 2 dimensions ont dès lors pour composantes

Iris setosa : $\bar{Z}_1 = -7.61$ et $\bar{Z}_2 = 0.215$
Iris versicolor : $\bar{Z}_1 = 1.83$ et $\bar{Z}_2 = -0.728$
Iris virginica : $\bar{Z}_1 = 5.78$ et $\bar{Z}_2 = 0.513$

Ces points sont représentés sur la figure ci-dessous qui illustre la position des trois groupes dans l'espace à 4 dimensions.



8.12 Classement dans plusieurs groupes

Soient à nouveau g groupes $H_1, \ldots, H_g$ de probabilité a priori $\pi_1, \ldots, \pi_g$ avec la convention $(\pi_1 + \ldots + \pi_g = 1)$ et un vecteur $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$. On dispose aussi d'une matrice d'observations $\underset{\sim}{X}_{n_i \times p}$ de chaque groupe $(i = 1, \ldots, g)$ avec $n = n_1 + \ldots + n_g$.

Soient $\bar{x}_1, \ldots, \bar{x}_g$ les vecteurs moyens et $\underset{\sim}{S}$ la matrice pondérée.

Dès lors, pour classer un sujet $\underset{\sim}{x}$ dans l'un des g groupes, on a besoin de $g-1$ fonctions discriminantes ; celles-ci s'écrivent, en prenant le groupe H_g comme référence

$$\begin{aligned} L_1(\underset{\sim}{x}) &= (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_g)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \tfrac{1}{2}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_g\right)\right] \\ L_2(\underset{\sim}{x}) &= (\underset{\sim}{\bar{x}}_2 - \underset{\sim}{\bar{x}}_g)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \tfrac{1}{2}\left(\underset{\sim}{\bar{x}}_2 + \underset{\sim}{\bar{x}}_g\right)\right] \\ &\ldots \\ L_{g-1}(\underset{\sim}{x}) &= (\underset{\sim}{\bar{x}}_{g-1} - \underset{\sim}{\bar{x}}_g)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \tfrac{1}{2}\left(\underset{\sim}{\bar{x}}_{g-1} + \underset{\sim}{\bar{x}}_g\right)\right] \end{aligned} \tag{8.25}$$

Lorsque $g = 2$, on retrouve l'expression (8.7).
On calcule ensuite les probabilités a posteriori qui s'écrivent

$$\begin{aligned} P_1(\underset{\sim}{x}) &= \frac{e^{\log \frac{\pi_1}{\pi_g} + L_1(\underset{\sim}{x})}}{1 + \sum_{j=1}^{g-1} e^{\log \frac{\pi_j}{\pi_g} + L_j(\underset{\sim}{x})}} \\ &\dots \\ P_{g-1}(\underset{\sim}{x}) &= \frac{e^{\log \frac{\pi_{g-1}}{\pi_g} + L_{g-1}(\underset{\sim}{x})}}{1 + \sum_{j=1}^{g-1} e^{\log \frac{\pi_j}{\pi_g} + L_j(\underset{\sim}{x})}} \\ P_g(\underset{\sim}{x}) &= 1 - P_1(\underset{\sim}{x}) - \ldots - P_{g-1}(\underset{\sim}{x}) \end{aligned} \tag{8.26}$$

Lorsque $g = 2$, on retrouve les équations (8.11).
La règle de classement s'énonce alors

"Classer le sujet $\underset{\sim}{x}$ dans le groupe pour lequel la probabilité a posteriori est maximale, c'est-à-dire classer dans H_j si $P_j(\underset{\sim}{x}) = \max\left\{P_1(\underset{\sim}{x}), \ldots, P_g(\underset{\sim}{x})\right\}$" (8.27)

On peut apprécier la qualité du classement en calculant la matrice de classement où l'on croise le groupe réel d'appartenance des observations avec le groupe dans lequel la règle de classement les classe.

	Groupe original			
Classement	H_1	H_2	$\ldots$	H_g
H_1	r_{11}	r_{12}	$\ldots$	r_{1g}
H_2	r_{21}	r_{22}	$\ldots$	r_{2g}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
H_g	r_{g1}	r_{g2}	$\ldots$	r_{gg}
Total	n_1	n_2	$\ldots$	n_g

Le taux de "classement correct" est égal à

$$C = \frac{r_{11} + r_{22} + \ldots + r_{gg}}{n_1 + n_2 + \ldots + n_g} \tag{8.28}$$

et le taux d'erreur total par resubstitution vaut $\varepsilon_R = 1 - C$.

On peut utiliser sans difficulté des techniques de sélection de variables comme on l'a fait précédemment afin de sélectionner les meilleures variables discriminantes.

Le problème d'analyse discriminante ou de classement multi-groupes peut aussi se résoudre par une méthode logistique généralisée mais peu de programmes existent pour appliquer cette méthode en pratique. On obtient $g-1$ fonctions discriminantes

$$\begin{array}{l} \Lambda_1(\underset{\sim}{x}) = b_{01} + \underset{\sim}{b}_1^T \underset{\sim}{x} \\ \ldots \qquad \ldots \qquad \ldots \\ \Lambda_{g-1}(\underset{\sim}{x}) = b_{0,g-1} + \underset{\sim}{b}_{g-1}^T \underset{\sim}{x} \end{array} \tag{8.29}$$

mais il convient d'appliquer une correction aux termes indépendants

$$b_{0,i}^* = b_{0,i} + \log \frac{n_g \pi_i}{n_i \pi_g} \qquad (i = 1, \ldots, g-1) \tag{8.30}$$

comme on l'a fait pour le cas $g = 2$ (voir (8.17)).

Les probabilités a posteriori s'écrivent alors

$$\begin{array}{l} P(H_i|\underset{\sim}{x}) = \dfrac{e^{\Lambda_i^*(\underset{\sim}{x})}}{1 + \sum_{j=1}^{g-1} e^{\Lambda_j^*(\underset{\sim}{x})}} \qquad (i = 1, \ldots, g-1) \\ \\ P(H_g|\underset{\sim}{x}) = 1 - \sum_{i=1}^{g-1} P(H_i|\underset{\sim}{x}) \end{array} \tag{8.31}$$

où $\Lambda_i^*(\underset{\sim}{x}) = b_{0,i}^* + \underset{\sim}{b}_i^T \underset{\sim}{x} \quad (i = 1, \ldots, g-1)$.

Remarque.

1. L'analyse discriminante multi-groupe peut s'assimiler à un problème de régression dans lequel la variable "dépendante" Y est qualitative à g modalités. On se rend compte qu'il n'est pas possible de résoudre ce problème avec une seule combinaison linéaire des variables indépendantes $X_1, \ldots, X_p$. En effet, les modalités de Y n'ont pas d'ordre et ne peuvent être numérotées de façon univoque. On l'a vu au Chapitre 2, $g-1$ variables binaires sont nécessaires. On comprend dès lors pourquoi on a besoin de $g-1$ fonctions discriminantes.

2. Dans les logiciels statistiques, on n'obtient pas directemment les coefficients des fonctions discriminantes. En fait, le programme donne le vecteur des coefficients des "scores discriminants", c'est-à-dire les expressions

$$\underset{\sim}{d}_i^T = \underset{\sim}{\bar{x}}_i^T \underset{\sim}{S}^{-1} \quad (i = 1, \ldots, g)$$

ainsi que les termes indépendants (ou constantes)

$$d_{0i} = \frac{1}{2}\underset{\sim}{d}_i^T \underset{\sim}{\bar{x}}_i = \frac{1}{2}\underset{\sim}{\bar{x}}_i^T \underset{\sim}{S}^{-1} \underset{\sim}{\bar{x}}_i \quad (i = 1, \ldots, g)$$

Dès lors, on voit immédiatement que les expressions (8.25) peuvent s'écrire

$$L_i(\underset{\sim}{x}) = (\underset{\sim}{d}_i - \underset{\sim}{d}_g)^T \underset{\sim}{x} - (d_{0i} - d_{0g}) \quad (i = 1, \ldots, g)$$

Il suffit donc de calculer les différences entre les scores discriminants et les constantes pour obtenir les fonctions discriminantes. Cette façon de procéder présente l'avantage de choisir à loisir le groupe de reférence.

En se servant des vecteurs moyens de chaque groupe de fleurs et de la matrice de variances - covariances pondérée donnée précédemment, on peut calculer la distance de Mahalanobis entre les trois groupes pris deux à deux et on obtient les valeurs suivantes, toutes hautement significatives ($p<0.0001$), confirmant la discrimination entre les trois groupes de fleurs.

	Setosa	Versicolor	Virginica
Setosa	0		
Versicolor	89.86	0	
Virginica	179.4	17.20	0

En supposant que les probabilités a priori sont toutes égales $\pi_1 = \pi_2 = \pi_3$, les scores discriminants sont donnés par les expressions suivantes

Variable	d_{Setosa}	$d_{Versicolor}$	$d_{Virginica}$
Constante	-85.21	-71.75	-103.27
Longueur sépale	2.354	1.570	1.245
Largeur sépale	2.359	0.707	0.369
Longueur pétale	-1.643	0.521	1.277
Largeur pétale	-1.734	0.643	2.108

En prenant le groupe 3 comme référence, les deux fonctions discriminantes s'écrivent

$$\begin{aligned} L_1(\underset{\sim}{x}) &= 18.06 + 1.11X_1 + 1.99X_2 - 2.92X_3 - 3.85X_4 \\ L_2(\underset{\sim}{x}) &= 31.52 + 0.325X_1 + 0.339X_2 - 0.756X_3 - 1.465X_4. \end{aligned}$$

Une fleur est classée dans H_1 (iris setosa) si $L_1(\underset{\sim}{x}) \geq L_2(\underset{\sim}{x})$ et $L_1(\underset{\sim}{x}) \geq 0$, dans H_2 (iris versicolor) si $L_2(\underset{\sim}{x}) > L_1(\underset{\sim}{x})$ et $L_2(\underset{\sim}{x}) \geq 0$, dans H_3 (iris virginica) sinon, soit si $L_1(\underset{\sim}{x}) < 0$ et $L_2(\underset{\sim}{x}) < 0$.

La matrice de classement obtenue par resubstitution s'écrit

	Groupe original			
Classement	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
Versicolor	0	48	1	49
Virginica	0	2	49	51
Total	50	50	50	150

On constate que le groupe iris setosa est complètement séparé des deux autres groupes qui se recouvrent partiellement comme on l'a vu précédemment. Le taux total de classement correct est égale à $C = (50+48+49)/150 = 0.98$. Dès lors, le taux d'erreur vaut $\varepsilon_r = 0.02$ (2%). La méthode d'extraction-réinsertion conduit à la même valeur.

Annexe I - Iris de Fisher

Setosa						Versicolor						Virginica					
N°	Y	X_1	X_2	X_3	X_4	N°	Y	X_1	X_2	X_3	X_4	N°	Y	X_1	X_2	X_3	X_4
1	1	50	33	14	2	1	2	65	28	46	15	1	3	64	28	56	22
2	1	46	34	14	3	2	2	62	22	45	15	2	3	67	31	56	24
3	1	46	36	10	2	3	2	59	32	48	18	3	3	63	28	51	15
4	1	51	33	17	5	4	2	61	30	46	14	4	3	69	31	51	23
5	1	55	35	13	2	5	2	60	27	51	16	5	3	65	30	52	20
6	1	48	31	16	2	6	2	56	25	39	11	6	3	65	30	55	18
7	1	52	34	14	2	7	2	57	28	45	13	7	3	58	27	51	19
8	1	49	36	14	1	8	2	63	33	47	16	8	3	68	32	59	23
9	1	44	32	13	2	9	2	70	32	47	14	9	3	62	34	54	23
10	1	50	35	16	6	10	2	64	32	45	15	10	3	77	38	67	22
11	1	44	30	13	2	11	2	61	28	40	13	11	3	67	33	57	25
12	1	47	32	16	2	12	2	55	24	38	11	12	3	76	30	66	21
13	1	48	30	14	3	13	2	54	30	45	15	13	3	49	25	45	17
14	1	51	38	16	2	14	2	58	26	40	12	14	3	67	30	52	23
15	1	48	34	19	2	15	2	55	26	44	12	15	3	59	30	51	18
16	1	50	30	16	2	16	2	50	23	33	10	16	3	63	25	50	19
17	1	50	32	12	2	17	2	67	31	44	14	17	3	64	32	53	23
18	1	43	30	11	1	18	2	56	30	45	15	18	3	79	38	64	20
19	1	58	40	12	2	19	2	58	27	41	10	19	3	67	33	57	21
20	1	51	38	19	4	20	2	60	29	45	15	20	3	77	28	67	20
21	1	49	30	14	2	21	2	57	26	35	10	21	3	63	27	49	18
22	1	51	35	14	2	22	2	57	29	42	13	22	3	72	32	60	18
23	1	50	34	16	4	23	2	49	24	33	10	23	3	61	30	49	18
24	1	46	32	14	2	24	2	56	27	42	13	24	3	61	26	56	14
25	1	57	44	15	4	25	2	57	30	42	12	25	3	64	28	56	21
26	1	50	36	14	2	26	2	66	29	46	13	26	3	62	28	48	18
27	1	54	34	15	4	27	2	52	27	39	14	27	3	77	30	61	23
28	1	52	41	15	1	28	2	60	34	45	16	28	3	63	34	56	24
29	1	55	42	14	2	29	2	50	20	35	10	29	3	58	27	51	19
30	1	49	31	15	2	30	2	55	24	37	10	30	3	72	30	58	16
31	1	54	39	17	4	31	2	58	27	39	12	31	3	71	30	59	21
32	1	50	34	15	2	32	2	62	29	43	13	32	3	64	31	55	18
33	1	44	29	14	2	33	2	59	30	42	15	33	3	60	30	48	18
34	1	47	32	13	2	34	2	60	22	40	10	34	3	63	29	56	18
35	1	46	31	15	2	35	2	67	31	47	15	35	3	77	26	69	23
36	1	51	34	15	2	36	2	63	23	44	13	36	3	60	22	50	15
37	1	50	35	13	3	37	2	56	30	41	13	37	3	69	32	57	23
38	1	49	31	15	1	38	2	63	25	49	15	38	3	74	28	61	19
39	1	54	37	15	2	39	2	61	28	47	12	39	3	56	28	49	20
40	1	54	39	13	4	40	2	64	29	43	13	40	3	73	29	63	18
41	1	51	35	14	3	41	2	51	25	30	11	41	3	67	25	58	18
42	1	48	34	16	2	42	2	57	28	41	13	42	3	65	30	58	22
43	1	48	30	14	1	43	2	61	29	47	14	43	3	69	31	54	21
44	1	45	23	13	3	44	2	56	29	36	13	44	3	72	36	61	25
45	1	57	38	17	3	45	2	69	31	49	15	45	3	65	32	51	20
46	1	51	38	15	3	46	2	55	25	40	13	46	3	64	27	53	19
47	1	54	34	17	2	47	2	55	23	40	13	47	3	68	30	55	21
48	1	51	37	15	4	48	2	66	30	44	14	48	3	57	25	50	20
49	1	52	35	15	2	49	2	68	28	48	14	49	3	58	28	51	24
50	1	53	37	15	2	50	2	67	30	50	17	50	3	63	33	60	25

X_1 = Longueur des sépales (mm) X_3 = Longueur des pétales (mm) Y= Groupe (1=iris setosa
X_2 = Largeur des sépales (mm) X_4 = Largeur des pétales (mm) 2=iris versicolor, 3=iris virginica)

Annexe II - Traumatisés crâniens

Issue à 6 mois (1=bonne récupération ou incapacité légère, 2=incapacité sévère ou état végétatif persistant, 3=décès), âge (années), taux de l'isoenzyme CK-BB (UI/l) dans le liquide céphalo-rachidien chez 60 traumatisés crâniens. (Source : Hans et al., 1985)

NPAT	Issue	Age	CK-BB	NPAT	Issue	Age	CK-BB
1	1	19	100	31	3	59	76
2	1	11	220	32	3	61	303
3	1	38	6	33	3	45	1560
4	1	7	281	34	3	61	353
5	1	17	17	35	3	20	1370
6	1	19	27	36	3	24	671
7	1	12	96	37	3	22	60
8	1	8	23	38	3	30	356
9	1	24	253	39	3	20	543
10	1	16	60	40	3	45	120
11	1	18	126	41	3	16	700
12	1	12	100	42	3	16	16
13	1	8	200	43	3	50	216
14	1	28	70	44	3	16	800
15	1	10	146	45	3	19	90
16	1	46	46	46	3	19	303
17	1	35	40	47	3	11	183
18	1	6	136	48	3	18	740
19	1	6	286	49	3	15	1256
20	2	8	230	50	3	56	523
21	2	23	509	51	3	40	350
22	2	19	283	52	3	18	126
23	2	4	140	53	3	18	153
24	2	29	80	54	3	20	913
25	2	23	576	55	3	19	193
26	2	20	76	56	3	41	323
27	2	62	206	57	3	51	443
28	3	17	253	58	3	29	156
29	3	29	490	59	3	21	463
30	3	7	1087	60	3	20	230

Annexe III - Patients avec cancer rectal

Survie (mois), dose de radiothérapie pré-opératoire ($0 = < 5000$ rads, $1 = \geq 5000$ rads), âge (années), sexe ($0 = femme,\ 1 = homme$) de 56 patients atteints d'un cancer rectal. (Source : Harris et Albert, 1991)

NPAT	Survie	Dose	Age	Sexe	NPAT	Survie	Dose	Age	Sexe
1	7	0	68	0	29	19	1	60	0
2	9	0	69	0	30	23*	1	54	0
3	12	0	68	0	31	24*	1	62	1
4	12	0	71	0	32	25*	1	55	0
5	19	0	77	1	33	26*	1	39	1
6	23	0	70	0	34	27	1	50	0
7	24	0	67	0	35	29*	1	58	1
8	24	0	68	1	36	30*	1	75	1
9	24	0	88	1	37	32*	1	61	1
10	24	0	89	1	38	33*	1	53	0
11	29*	0	28	1	39	33*	1	57	1
12	34	0	73	1	40	35	1	50	1
13	41	0	60	0	41	35	1	78	1
14	54	0	60	0	42	35*	1	55	1
15	72*	0	44	1	43	35*	1	65	1
16	78	0	82	0	44	35*	1	73	1
17	80*	0	62	0	45	36	1	53	0
18	83*	0	53	0	46	38*	1	47	1
19	92*	0	66	0	47	51*	1	60	1
20	139*	0	63	0	48	54*	1	54	1
21	139*	0	68	1	49	57	1	66	1
22	9	1	77	0	50	60*	1	64	0
23	12	1	55	0	51	67	1	60	1
24	12*	1	78	0	52	70	1	41	1
25	13*	1	47	0	53	87*	1	58	1
26	14*	1	69	1	54	89*	1	45	1
27	16	1	68	0	55	98*	1	73	1
28	18*	1	62	1	56	120*	1	63	1

* Durée de vie censurée

Annexe IV - Tables statistiques

Table A Loi Normale $Z \sim N(0,1)$.
Probabilités supérieures $P[Z > z]$, $z \geq 0$

Table B Loi Chi-carré à ν degrés de liberté χ^2_ν
Quantiles supérieurs $Q_{\chi^2}(1-\alpha;\nu)$

Table C Loi t de Student à ν degrés de liberté t_ν
Quantiles supérieurs $Q_t(1-\frac{\alpha}{2};\nu)$
Note : la ligne $\nu = \infty$ correspond aux quantiles supérieurs gaussiens $Q_Z(1-\frac{\alpha}{2})$

Table D Loi F de Snedecor à ν_1 et ν_2 degrés de liberté F_{ν_1,ν_2}
Quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ pour $\alpha = 0.05$ et $\alpha = 0.01$

Table A Loi Normale $Z \sim N(0,1)$
Probabilités supérieures $P[Z > z]$, $z \geq 0$

z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002

Table B Loi Chi-carré à ν degrés de liberté χ^2_ν
Quantiles supérieurs $Q_{\chi^2}(1-\alpha;\nu)$

	$1-\alpha$					
ν	0.90	0.95	0.975	0.99	0.995	0.999
1	2.706	3.841	5.024	6.635	7.879	10.83
2	4.605	5.991	7.378	9.210	10.60	13.82
3	6.251	7.815	9.348	11.34	12.84	16.27
4	7.779	9.488	11.14	13.28	14.86	18.47
5	9.236	11.07	12.83	15.09	16.75	20.52
6	10.64	12.59	14.45	16.81	18.55	22.46
7	12.02	14.07	16.01	18.48	20.28	24.32
8	13.36	15.51	17.53	20.09	21.95	26.12
9	14.68	16.92	19.02	21.67	23.59	27.88
10	15.99	18.31	20.48	23.21	25.19	29.59
11	17.28	19.68	21.92	24.72	26.76	31.26
12	18.55	21.03	23.34	26.22	28.30	32.91
13	19.81	22.36	24.74	27.69	29.82	34.53
14	21.06	23.68	26.12	29.14	31.32	36.12
15	22.31	25.00	27.49	30.58	32.80	37.70
16	23.54	26.30	28.85	32.00	34.27	39.25
17	24.77	27.59	30.19	33.41	35.72	40.79
18	25.99	28.87	31.53	34.81	37.16	42.31
19	27.20	30.14	32.85	36.19	38.58	43.82
20	28.41	31.41	34.17	37.57	40.00	45.31
21	29.62	32.67	35.48	38.93	41.40	46.80
22	30.81	33.92	36.78	40.29	42.80	48.27
23	32.01	35.17	38.08	41.64	44.18	49.73
24	33.20	36.42	39.36	42.98	45.56	51.18
25	34.38	37.65	40.65	44.31	46.93	52.62
26	35.56	38.89	41.92	45.64	48.29	54.05
27	36.74	40.11	43.19	46.96	49.64	55.48
28	37.92	41.34	44.46	48.28	50.99	56.89
29	39.09	42.56	45.72	49.59	52.34	58.30
30	40.26	43.77	46.98	50.89	53.67	59.70
35	46.06	49.80	53.20	57.34	60.27	66.62
40	51.81	55.76	59.34	63.69	66.77	73.40
45	57.51	61.66	65.41	69.96	73.17	80.08
50	63.17	67.50	71.42	76.15	79.49	86.66
60	74.40	79.08	83.30	88.38	91.95	99.61
70	85.53	90.53	95.02	100.4	104.2	112.3
80	96.58	101.9	106.6	112.3	116.3	124.8
90	107.6	113.1	118.1	124.1	128.3	137.2
100	118.5	124.3	129.6	135.8	140.2	149.4

Table C Loi t de Student à ν degrés de liberté t_ν
Quantiles supérieurs $Q_t(1-\frac{\alpha}{2};\nu)$

	$1-\frac{\alpha}{2}$					
ν	0.90	0.95	0.975	0.99	0.995	0.999
1	3.0777	6.3138	12.706	31.821	63.657	318.31
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.327
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.215
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853
8	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247
12	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296
13	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520
14	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874
15	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328
16	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852
35	1.3062	1.6896	2.0301	2.4377	2.7238	3.3400
40	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069
45	1.3006	1.6794	2.0141	2.4121	2.6896	3.2815
50	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317
70	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108
80	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953
90	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737
120	1.2886	1.6577	1.9799	2.3578	2.6174	3.1595
200	1.2858	1.6525	1.9719	2.3451	2.6006	3.1315
∞	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902

Table D Loi de Snedecor à ν_1 et ν_2 degrés de liberté
Quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ pour $1-\alpha = 0.95$

					ν_1					
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.02	1.83	1.61	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Table D Loi de Snedecor à ν_1 et ν_2 degrés de liberté
Quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ pour $1-\alpha = 0.99$

	ν_1									
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.61
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80	3.43	3.01
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.76
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	2.66
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Table des matières