

I. Analyse discriminante (deux groupes)

1. Coefficients discriminants de Fisher

L'objectif poursuivi est de pouvoir classer des individus entre deux groupes codés par une variable binaire. Pour cela, on dispose de plusieurs variables explicatives. On va chercher une combinaison linéaire de ces variables de la forme $DF = \mathbf{b}_1 X_1 + \mathbf{b}_2 X_2 + \dots + \mathbf{b}_p X_p$ (ou $DF = X \beta$ sous forme matricielle) qui maximise la distinction entre les groupes. On va en d'autres termes chercher β tel que la variation entre les groupes soit la plus grande possible et la variation au sein des groupes, la plus petite possible.

- Evaluation de la variation au sein des groupes : la variation au sein de chaque groupe est la somme des différences au carré entre chaque observation et la moyenne de son groupe.

Soit $X1_{n,p}$, la matrice des données du groupe 1 (n individus et p variables) et $X2_{n,p}$, celle du groupe 2, n_1 le nombre d'individus du groupe 1 et n_2 , le nombre d'individus du groupe 2 :

- la moyenne du groupe 1 est : $\bar{X}1_p = \frac{1}{n_1} \sum_{i=1}^{n_1} X1_{ip}$ et $\bar{X}1 = \begin{bmatrix} \bar{X}1_1 \\ \vdots \\ \bar{X}1_p \end{bmatrix}$ est le vecteur

des moyennes.

- la somme des carrés et des produits croisés entre les variables au sein du groupe 1 est :

$$W1_{xy} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X1_{ix} - \bar{X}1_x)(X1_{iy} - \bar{X}1_y)$$

- on obtient une estimation de la variation globale au sein des groupes en addition les matrices WSSCP de chaque groupe : $WSSCP = W1 + W2$. (SSCP signifie sum of square and cross product).

- Evaluation de la variation entre les groupes : la variation entre les groupes est la différences entre la moyenne des groupes et la moyenne globale de la population.

- la moyenne global est la moyenne des groupes pondérées le nombre

$$\text{d'individus : } \bar{X}_p = \frac{1}{n} \sum_{k=1}^2 n_k \bar{X}_{kp}, \text{ où } n = \sum_{k=1}^2 n_k .$$

- la BSSCP entre les groupes est obtenue par :

$$BSSCP_{xy} = \sum_{k=1}^2 n_k (\bar{X}_{kx} - \bar{X}_x)(\bar{X}_{ky} - \bar{X}_y)$$

On cherche donc un vecteur des paramètres β qui maximise le rapport :

$$\frac{\mathbf{b}'\mathbf{BSSCP}\mathbf{b}}{\mathbf{b}'\mathbf{WSSCP}\mathbf{b}}$$

La solution est la suivante : $\mathbf{b} = (n_1 + n_2 - 2)\mathbf{WSSCP}^{-1}[\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2]$

Si l'on désire tester le niveau de signification de la discrimination obtenue, cela revient à tester l'hypothèse nulle :

H_0 : moyenne du groupe 1 = moyenne du groupe 2 (ce qui est équivalent à $\beta = 0$)

Un statistique possible pour ce test est le Wilks Lambda. Sachant que $\mathbf{TSSCP} = \mathbf{BSSCP} + \mathbf{WSSCP}$, le Wilks Lambda est le rapport :

$$\frac{|\mathbf{WSSCP}|}{|\mathbf{TSSCP}|}$$

2. Application à la classification

Pour déterminer le "cutoff point" entre les groupes, on procède comme suit :

- on calcule pour chaque individu de chaque groupe les valeur de DF.
- on calcule la valeur moyenne de DF pour chaque groupe.
- le "cutoff point" à la moyenne arithmétique des DF moyens de chaque groupe.

On peut également tenir compte de la probabilité a priori qu'une observation appartienne à un groupe. Pour estimer la probabilité a priori qu'un individu donné appartienne au groupe 1, on utilise alors le rapport :

$$\text{Prob1} = \frac{n_1}{n_1 + n_2}$$

et on utilise comme cutoff : $\text{Cutoff} = \text{Cutoff} + \ln\left(\frac{1 - \text{Prob1}}{\text{Prob1}}\right)$.