# Projet "fil rouge"

Les exploitations agricoles (articles de D. Dubois)

Partie (1) : Realiser une ACP, appliquez la méthodologie et les méthodes d'interpretation vues en cours

Partie (2) : Utilisation des resultats de l'ACP pour definir une fonction score et une regle de decision pour prédire si une exploitation sera "saine" ou "defaillante", connaissant ses 22 ratios financiers

2.1) A partir de la première composante principale $\psi^1$,

- Faire le diagramme boxplot de $\psi^1$ pour les exploitations "saine" et les exploitations "defaillantes"
- Calculer la moyenne de $\psi^1$ dans ces deux groupes, et verifier que la moyenne pondérée de ces deux moyenne est 0.

2.2) Une règle de classement géométrique pour une exploitation $i$ appartenant à l'échantillon consiste à la positionner par rapport au point pivot qui est le milieu du segment des deux moyennes $\mu_1 = -2,164$ et $\mu_2 = 2,328$ soit $\frac{\mu_1+\mu_2}{2} = 0,082$ :

si $\psi_i^1 < 0,082$, alors l'exploitation $i$ est déclarée "saine"

si $\psi_i^1 > 0,082$, alors l'exploitation $i$ est déclarée "defaillante"

Calculer à partir de cette règle de classement une nouvelle variable "PRED" qui vaudra "Sain" ou "défaillant" selon la règle précédente.

Vérifiez que vous retrouvez le tableau de contingence suivant :

| DIFF \ PRED | Sain | Défaillant |
|---|---|---|
| Sain | 599 | 54 |
| Défaillant | 124 | 483 |

En déduire les pourcentages intéressants et interpréter.

En particulier, on trouve un taux de mal classés de 14,12 %

2.3) On veut maintenant pouvoir utiliser cette règle de classement pour pouvoir prédire la classe d'une nouvelle exploitation $i_0$, dont on connait les caractéristiques financières. On sait que $\psi_i^1 = Z\nu = \sum_{j=1}^{p} \nu_j z_{ij}$

et $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ $\Rightarrow \psi_i^1 = \sum_{j=1}^{p} \frac{\nu_j}{s_j} x_{ij} - \sum_{j=1}^{p} \nu_j \frac{\bar{x}_j}{s_j}$

$\frac{\nu_j}{s_j}$ : coef. j ; $\sum_{j=1}^{p} \nu_j \frac{\bar{x}_j}{s_j}$ : constante

On en déduit la fonction score suivante

$$S(i) = \sum_{j=1}^{p} a_j x_{ij} + \text{constante}$$

Vérifiez : - que vous retrouvez ce tableau

| | Fonction S |
|---|---|
| $R_1$ | 0,76 |
| $R_2$ | -0,83 |
| ⋮ | |
| $R_{37}$ | -0,12 |
| constant | -1,53 |

- qu'en appliquant cette fonction à toutes les exploitations de l'échantillon vous retrouvez $\psi^1$.

2.4) On sait que le taux d'erreur calculé sur l'échantillon d'apprentissage (appelé taux apparent d'erreur) calculé au 2.2) sous estime souvent le taux d'erreur ne que l'on utilise les mêmes observations pour le calculer que celles qui ont servi à trouver la règle géométrique de classement. On utilise donc souvent la méthode de l'échantillon-test pour estimer ce taux d'erreur: Cette méthode consiste à partager en deux parties l'échantillon:

- une partie sert d'échantillon d'apprentissage de la règle de décision
- l'autre partie sert d'échantillon test et permet de tester la règle d'affectation et donc de calculer le taux d'erreur

2.4.1) Calculer ce taux d'erreur en prenant $2/3$ des observations dans l'échantillon d'apprentissage (les 840 premières exploitations) et $1/3$ dans l'échantillon test (les 420 dernières)

2.4.2) Calculer ce taux d'erreur en tirant cette fois aléatoirement (en vous aidant de la fonction sample()) un échantillon d'apprentissage (de 840 exploitations) et un échantillon test (des 240 exploitations). Recommencez 5 fois et calculez le taux d'erreur moyen. Commentez.

2.4.3) Expliquez ce dernier résultat en vous aidant de la remarque suivante: "Le taux d'erreur apparent est d'autant plus faible que le modèle est complexe (sur-paramétrisation". Proposez une "solution".