Statistique
Informatique
Mathématique
appliquées

Septembre 2008

gembloux
faculté universitaire
des sciences agronomiques

MACRO MINITAB POUR LA VERIFICATION DE L'INTERET D'UNE REGLE DE DECISION EN ANALYSE DISCRIMINANTE

R. PALM*

1. Introduction

L'analyse discriminante décisionnelle regroupe un ensemble de méthodes dont l'objectif est de définir une règle de classement d'individus dans des groupes préalablement définis. L'affectation de l'individu à l'un de ces *g* groupes se fait sur base de *p* caractéristiques, c'est-à-dire de *p* variables, observées sur cet individu et la règle de classement est établie en fonction de ces mêmes *p* caractéristiques, observées sur des échantillons provenant des *g* groupes.

A l'issue de l'établissement d'une règle d'affectation, le praticien peut se demander si cette règle classe mieux les individus dans les groupes que ne le ferait le hasard. Différentes solutions sont proposées dans la littérature et deux approches ont été retenues. Une macro Minitab reprenant ces deux approches est proposée aux utilisateurs de ce logiciel. Elle est disponible sur le site de l'Unité de Statistique, Informatique et Mathématique appliquées de la FUSAGx[1] à l'adresse suivante :

www.fsagx.ac.be/si/

en cliquant sur le lien Macros, et, ensuite, sur la rubrique en question.

Nous présentons d'abord le tableau de classement (paragraphe 2). Ensuite, nous examinons un critère basé sur l'affectation aléatoire des individus dans les groupes (paragraphe 3) et un critère basé sur l'affectation de tous les individus dans le groupe de probabilité *a priori* maximale (paragraphe 4). Nous présentons alors la macro EFFICDISCRI (paragraphe 5), avant de conclure (paragraphe 6).

* Professeur à la Faculté universitaire des Sciences agronomiques de Gembloux.
1. Faculté universitaire des Sciences agronomiques de Gembloux (Belgique)

2. Tableau de classement

Un des outils utilisés en analyse discriminante pour évaluer la performance d'une règle de classement est le tableau de classement, qui reprend, en fonction des groupes d'origine, les groupes dans lesquels la règle classe *n* individus d'origine connue.

A titre d'illustration, nous reprenons l'exemple proposé par ALBERT et HARRIS [1987] concernant 218 patients répartis en quatre groupes en fonction de la maladie du foie dont ils souffrent, et pour lesquels on dispose des teneurs du sang en trois enzymes. Cet exemple a été repris par PALM [2008] pour illustrer l'utilisation de la macro DLOGISTIC, qui réalise l'analyse discriminante logistique. La note qui décrit cette macro ainsi que la macro elle-même sont disponibles sur le site de l'Unité, à l'adresse donnée dans l'introduction.

Le tableau 1 donne les résultats du reclassement des 218 patients à partir de la règle d'affectation obtenue par l'analyse discriminante logistique, en considérant que les probabilités *a priori* d'appartenance aux quatre groupes sont égales à 0,40, 0,20, 0,15 et 0,25.

Les effectifs repris sur la diagonale principale correspondent aux patients bien classés. Ainsi, par exemple, 53 patients atteints de la pathologie 1 sont effectivement affectés au groupe 1 par la règle. Par contre, les effectifs hors de la diagonale principale représentent les nombres de patients mal classés. Ainsi, pour les patients atteints de la pathologie 1, trois sont classés dans le groupe 2 et un est classé dans le groupe 1.

Tableau 1. Classement de patients en fonction de leur pathologie réelle et de leur groupe d'affectation.

Origine	Affectation				Totaux
	1	2	3	4	
1	53	3	1	0	57
2	3	38	1	2	44
3	2	2	21	15	40
4	0	1	8	68	77
Totaux	58	44	31	85	218

Le tableau 2 reprend la structure générale d'un tableau de classement et donne les notations qui seront utilisées par la suite. Les totaux des colonnes, qui correspondent aux nombres d'individus classés dans les différents groupes, n'ont pas été repris dans le tableau, car ils ne seront pas utilisés par la suite. Seul le total de la dernière colonne est utile : il correspond à l'effectif total :

$$n = \sum_{i=1}^{g} n_i .$$

Nous utiliserons également les probabilités *a priori*, qui correspondent aux proportions des groupes dans la population globale et nous les notons p_i (i = 1, ..., g).

Tableau 2. Structure générale d'un tableau de classement.

Origine	Affectation						Totaux
	1	2	...	*j*	...	*g*	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1g}	n_1
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2g}	n_2
⋮	⋮	⋮	...	⋮	...	⋮	⋮
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ig}	n_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
g	n_{g1}	n_{g2}	...	n_{gj}	...	n_{gg}	n_g

Le tableau de classement concerne toujours le classement d'individus d'origine connue et, dans la pratique, deux types de tableaux peuvent se présenter. Le tableau peut concerner le reclassement des individus qui sont utilisés pour la définition de la règle de classement, ou au contraire, d'autres individus. Dans le premier cas, on parle de resubstitution et, dans le second cas, de validation externe. Une variante du premier type est le tableau obtenu par validation croisée. Pour obtenir ce dernier tableau, la règle d'affectation est recalculée n fois sur $n - 1$ individus, en supprimant chaque fois un individu et en classant ensuite l'individu supprimé avec la règle ainsi calculée.

Les critères qui seront définis aux paragraphes 3 et 4 sont applicables aux différents tableaux définis ci-dessus, mais leur interprétation devra tenir compte de la nature du tableau analysé.

3. Critère basé sur l'affectation aléatoire

La question posée est la suivante : peut-on considérer que le nombre d'individus bien classés suite à l'utilisation de la règle d'affectation est significativement supérieur à ce qu'on obtiendrait en répartissant les individus au hasard dans les groupes, avec des probabilités égales aux probabilités *a priori* ?

L'analyse peut être réalisée pour un groupe d'origine donné, ou, au contraire, globalement pour l'ensemble des groupes.

Nous examinons d'abord le cas d'un groupe donné et nous prenons comme exemple le groupe 1 du tableau 1, qui comporte 57 individus.

Pour répartir au hasard ces 57 individus dans les quatre groupes avec des probabilités égales aux probabilités *a priori*, qui sont égales à 0,40, 0,20, 0,15 et 0,25, on pourrait procéder de la manière suivante, en utilisant une urne comportant 100 jetons numérotés de 1 à 100. Pour affecter au hasard un individu à un groupe, on prèlève un jeton dans l'urne et on affecte l'individu :

- au groupe 1 si le nombre lu sur le jeton est compris entre 1 et 40,
- au groupe 2 si le nombre lu sur le jeton est compris entre 41 et 60,
- au groupe 3 si le nombre lu sur le jeton est compris entre 61 et 75,
- au groupe 4 si le nombre lu sur le jeton est compris entre 76 et 100.

On répète 57 fois le procédé pour affecter les 57 individus, en remettant dans l'urne les jetons au fur et à mesure de leur tirage.

Le nombre d'individus bien classés, c'est-à-dire affectés au groupe 1, dépend des hasards du tirage et peut donc être décrit par une variable aléatoire, que nous désignons par X. Cette variable aléatoire X possède une distribution binomiale de paramètres :

$$n_0 = 57 \text{ et } p_0 = 0{,}4.$$

Si la règle de classement utilisée classe mieux les individus que la répartition aléatoire, on doit s'attendre à ce que la proportion d'individus bien classés soit supérieure à p_0 et, pour vérifier si la proportion p d'individus bien classés par la règle est significativement supérieure à p_0, on réalise le test de conformité d'une porportion. L'hypothèse nulle et l'hypothèse alternative s'écrivent :

$$H_0 : p = p_0 \; ; \; H_1 : p > p_0.$$

On rejette l'hypothèse nulle si :

$$P\,(X \geq 53 \mid n_0, p_0) < 0{,}05,$$

ce qui est le cas, car $P\,(X \geq 53) \approx 0{,}0000$.

Un raisonnement analogue peut être tenu pour les autres lignes du tableau et on peut conclure que, pour tous les groupes, la règle de classement donne une proportion d'individus bien classés significativement supérieure à la probabilité *a priori*.

D'une manière générale, pour un groupe donné, on teste l'hypothèse nulle :

$$H_0 : p = p_i \; ; \; H_1 : p > p_i,$$

en comparant la probabilité :

$$P\,(\, X \geq n_{ii} \mid n_i, p_i)$$

au niveau de signification α qu'on se fixe.

Si on s'intéresse à l'ensemble du tableau 1 et sous l'hypothèse d'une répartition aléatoire dans les groupes avec des probabilités égales aux probabilités *a priori* pour les différents groupes, la proportion d'individus bien classés sera égale, en moyenne, à :

$$[(57)(0{,}40) + (44)(0{,}20) + (40)\,(0{,}15) + (77)\,(0{,}25)] \,/\, 218 = 0{,}2608.$$

Le nombre total d'individus bien classés sera, dans ce cas, une réalisation d'une variable aléatoire binomiale de paramètres :

$$n_0 = 218 \text{ et } p_0 = 0{,}2608.$$

Le nombre d'individus bien classés par l'utilisation de la règle est de :

$$53 + 38 + 21 + 68 = 180,$$

soit une proportion de 0,8256.

On rejette donc l'hypothèse nulle :

$$H_0 : p = p_0$$

contre l'alternative :

$$H_1 : p > p_0$$

si :

$$P\,(X \geq 180 \mid n_0, p_0) < 0{,}05,$$

ce qui est le cas, $P\,(X \geq 180)$ étant pratiquement nul.

D'une manière générale, on a :

$$p_0 = \frac{1}{n}\sum_{i=1}^{g} p_i n_i \quad \text{et} \quad x = \sum_{i=1}^{g} n_{ii}.$$

On teste l'hypothèse nulle :

$$H_0 : p = p_0 \; ; \quad H_1 : p > p_{0,}$$

en comparant la probabilité :

$$P\,(X \geq x \mid n, p_0),$$

au niveau de signification α fixé.

Indépendamment des tests statistiques, on peut également chiffrer le gain dû à l'utilisation de la règle d'affectation, par rapport à l'affectation aléatoire. Pour le premier groupe, le taux de patients bien classés par la règle est de :

$$53 / 57 = 0{,}9298,$$

et le taux de bien classés lors de l'affectation aléatoire est de 0,40. La différence entre les deux taux est de :

$$0{,}9298 - 0{,}40 = 0{,}5298.$$

HUBERTY [1994] propose d'exprimer cette différence en proportion du taux d'individus mal classés lors de la répartition aléatoire :

$$I = \frac{0{,}9298 - 0{,}40}{1 - 0{,}40} = 0{,}8830.$$

On peut, de la même manière, calculer l'indice pour les autres groupes et aussi un indice global pour l'ensemble du tableau :

$$I = \frac{0{,}8257 - 0{,}2608}{1 - 0{,}2608} = 0{,}7642.$$

On conclut donc que, par rapport à l'affectation aléatoire, l'utilisation de la règle permet de réduire la proportion de mal classés de 88 % pour le premier groupe et de 76 % pour l'ensemble des groupes.

4. Critère basé sur la probabilité *a priori* maximale

Dans le paragraphe précédent, on compare une proportion d'individus bien classés par la règle d'affectation à la proportion d'individus bien classés par une affectation aléatoire, qui tient cependant compte des probabilités *a priori*. Ainsi, pour l'ensemble des groupes, la proportion de patients bien classés est de : 0,8257 lors de l'utilisation de la règle et de 0,2608 pour l'affectation aléatoire.

Un taux de classement correct plus important que celui donné par le hasard peut cependant être obtenu, dans certains cas, par une affectation de tous les individus au groupe pour lequel la probabilité *a priori* est la plus importante. Pour l'exemple, la probabilité *a priori* est la plus grande pour le groupe 1 (p_1 = 0,40). En classant tous les individus dans ce groupe, le taux de classement correct serait de :

$$57 / 218 = 0{,}2615.$$

Il peut donc être utile de vérifier si la proportion d'individus bien classés est bien significativement supérieure à la proportion *a priori* maximum, c'est-à-dire si :

$$P\,(X \geq 180 \mid n_0, p_0) < 0{,}05,$$

X étant une variable binomiale de paramètre n_0 = 218 et p_0 = 0,2615. La probabilité étant pratiquement nulle, on peut conclure que la proportion d'individus bien classés par l'utilisation de la règle est supérieure à la probabilité de bien classés lorsque tous les individus sont classés dans le groupe pour lequel la probabilité *a priori* est la plus grande.

De façon plus générale, en désignant par m le groupe ayant la probabilité *a priori* maximum, on a :

$$p_0 = n_m / n \quad \text{et} \quad x = \sum_{i=1}^{g} n_{ii},$$

et on rejette l'hypothèse nulle :

$$H_0 : p = p_0$$

contre l'alternative :

$$H_1 : p > p_0$$

si :

$$P\,(X \geq x \mid n, p_0) < \alpha.$$

On peut également quantifier le gain de porportion des individus bien classés par rapport au classement de l'ensemble des patients dans le groupe ayant la probabilité *a priori* maximum, en calculant l'indice défini au paragraphe précédent. On trouve :

$$I = \frac{0{,}8257 - 0{,}2615}{1 - 0{,}2615} = 0{,}7640.$$

Pour cet exemple, il n'y a pas de différence importante entre la proportion d'individus bien classés sur la base de la probabilité *a priori* maximum (p_0 = 0,2615) et la proportion de bien classés sur la base de l'affectation aléatoire des individus (p_0 = 0,2608 : paragraphe 2). Il en résulte que les tests basés sur ces deux modes de répartition sont fort équivalents, de même que les deux indices *I*. Il n'en est cependant pas toujours ainsi. A titre d'illustration, les calculs ont été réalisés pour le tableau 3, qui reprend des données fictives, en considérant que les probabilités *a priori* sont proportionnelles aux fréquences des groupes, soit p_1 = 0,2 et p_2 = 0,8. Le taux de classement correct pour la règle est de :

$$(20 + 107) / 150 = 0{,}85,$$

le taux de classement correct pour l'affectation aléatoire des individus est de :

$$[(0{,}2)\ (30) + (0{,}8)\ (120)] / 150 = 0{,}68$$

et le taux de classement correct pour l'affectation de tous les individus au groupe 2 est de 0,80.

Ces valeurs nous conduiraient à conclure que la règle donne un nombre d'individus bien classés significativement supérieur à la répartition aléatoire dans les groupes mais pas à l'affectation systématique des individus au groupe 2, la probabilité associée à ce dernier test étant de 0,09.

Tableau 3. Tableau de classement pour deux groupes (données fictives).

Origine	Affectation		Totaux
	1	2	
1	20	10	30
2	13	107	120

5. La macro EFFICDISCRI

A partir du tableau de classement et des probabilités *a priori*, la macro réalise les tests et calcule les indices qui ont été définis au paragraphe 3, ce qui permet de comparer les résultats de l'affectation des individus par la règle utilisée aux résultats de l'affectation aléatoire, tenant compte des probabilités *a priori*. En outre, si le taux de classement correct par l'affectation de tous les individus au groupe de probabilité *a priori* maximale est supérieur à celui obtenu par la répartition aléatoire, la macro compare les résultats obtenus par la règle utilisée à l'affectation de tous les individus au groupe de probabilité *a priori* maximale. Différentes options permettent d'enregistrer les résultats et de supprimer les impressions. Des informations complémentaires sont données dans la notice d'utilisation qui accompagne la macro.

La figure 1 reprend la commande et les résultats de l'exécution de la macro sur les données du tableau 1.

Les donnnées de départ sont préalablement enregistrées dans les six premières colonnes du fichier Minitab et sont reprises dans la première partie des résultats.

Le deuxième tableau donne, pour chaque groupe d'origine, le taux de classement correct pour le tableau initial, le taux de classement correct pour une affectation aléatoire, c'est-à-dire aussi la probabilité *a priori*, l'indice de réduction du taux de classement erroné par la prise en compte de la règle de classement par rapport à la répartition aléatoire et la probabilité liée au test de conformité de la proportion d'individus bien classés par la règle à la probabilité *a priori*. Les mêmes informations sont ensuite fournies pour l'ensemble des groupes.

Enfin, la dernière partie du tableau permet la comparaison de la règle de classement par rapport au classement de tous les individus dans le groupe présentant la probabilité *a priori* la plus grande.

On retrouve bien tous les résultats qui ont été obtenus aux paragraphes 3 et 4.

```
COMMANDES
---------

 %EFFICDISCRI c1 c2-c5;
     PRIORS c6.


RESULTATS
---------

Tableau de classement et probabilités a priori

Row  GROUPEOR  TABL.1  TABL.2  TABL.3  TABL.4    A_PRIORI
  1         1      53       3       1       0        0.40
  2         2       3      38       1       2        0.20
  3         3       2       2      21      15        0.15
  4         4       0       1       8      68        0.25

Comparaison par rapport à la répartition aléatoire

Row  GROUPEOR       TRG   TCG        IG      BPROB
  1         1  0.929825  0.40  0.883041  0.0000000
  2         2  0.863636  0.20  0.829545  0.0000000
  3         3  0.525000  0.15  0.441176  0.0000000
  4         4  0.883117  0.25  0.844156  0.0000000


TR        0.825688
TC        0.260780
I         0.764195
BProba    0

Comparaison par rapport à l'affectation
  au groupe de probabilité a priori maximum

TR       0.825688
TCMAX    0.261468
IMAX     0.763975
BPmax    0
```

Figure 1. Macro EFFICDISCRI : commande et résultats.

6. Conclusions

La macro EFFICDISCRI permet de compléter utilement les informations résultant d'une analyse discriminante linéaire ou logistique réalisée par la commande DISCRIMINANT de Minitab ou d'une analyse réalisée à l'aide de tout autre logiciel. Il suffit, à cet effet, d'encoder le tableau de classement et les probabilités *a priori*.

Cette macro a par ailleurs été intégrée à la macro DLOGISTIC, disponible sur le site de l'Unité à l'adresse donnée dans l'introduction, qui réalise l'analyse discriminante logistique.

Nous avons signalé, dans l'introduction, que le tableau de classement pouvait concerner soit les données utilisées pour l'établissement de la règle (resubstitution), soit des données supplémentaires (validation), soit les données qui ont été utilisées pour le calcul de la règle d'affectation mais en faisant appel à une procédure de validation croisée. L'interprétation des résultats doit évidemment tenir compte de la nature du tableau de départ. Ainsi, le fait de

conclure à la supériorité d'une règle d'affectation sur la répartition aléatoire à partir d'un tableau de resubstitution ne garantit pas que le taux d'individus bien classés restera supérieur lorsque la règle sera utilisée pour classer d'autres individus que ceux utilisés pour l'établissement de la règle.

Bibliographie

ALBERT A., HARRIS E. [1987]. *Multivariate interpretation of clinical laboratory data*. New York, Dekker, 312 p.

HUBERTY C. J. [1994]. *Applied discriminant analysis*. New York, Wiley, 466 p.

PALM R. [2008] Macro Minitab pour l'analyse discriminante logistique. <http://www.fsagx.ac.be/si/>