Statistique
Informatique
Mathématique
appliquées

Septembre 2008

gembloux
faculté universitaire
des sciences agronomiques

MACRO MINITAB POUR L'ANALYSE DISCRIMINANTE LOGISTIQUE

R. PALM*

1. Introduction

L'analyse discriminante décisionnelle a pour objectif de définir une règle permettant de classer un individu dans un groupe particulier parmi *g* groupes possibles. Cette affectation de l'individu à un groupe donné se fait sur la 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 ou populations.

Différentes méthodes ont été proposées pour l'établissement de cette règle d'affectation. La méthode la plus ancienne est l'analyse discriminante linéaire, proposée par FISHER [1936], qui s'applique aux situations où les variables décrivant les groupes sont des variables multinormales, ayant même matrice de variances et covariances pour les différents groupes. Une extension est l'analyse quadratique discriminante, qui s'applique à des populations normales ayant des matrices de variances et covariances différentes d'un groupe à l'autre.

Des informations concernant ces méthodes sont données, notamment, dans les livres de HUBERTY [1994] et MCLACHLAN [1993], ainsi que dans la note publiée par PALM [1999], disponible sur internet à l'adresse précisée ci-dessous.

Les hypothèses à la base de l'analyse discriminante linéaire ou quadratique sont assez restrictives, et lorsqu'elles ne sont pas vérifiées, les règles de décision qui en découlent conduisent à des taux de classement erronés plus grands que ceux liés à d'autres méthodes et notamment à l'analyse discriminante logistique.

* Professeur à la Faculté universitaire des Sciences agronomiques de Gembloux.

L'Unité de Statistique, Informatique et Mathématique appliquées de la FUSAGx[1] propose aux utilisateurs de Minitab une macro permettant de réaliser l'analyse discriminante logistique. Cette macro ainsi que sa notice d'utilisation sont disponibles à l'adresse suivante :

www.fsagx.ac.be/si/

en cliquant sur le lien Macros, puis sur le thème en question.

L'analyse discriminante logistique étant très étroitement liée à la régression logistique, nous rappelons d'abord quelques principes de cette régression, d'une part dans le cas de deux groupes (paragraphe 2) et, d'autre part, dans le cas de plus de deux groupes (paragraphe 3).

Ensuite, nous donnons quelques informations concernant les schémas d'échantillonnage utilisés lors de la collecte des données (paragraphe 4) et nous examinons comment les résultats de la régression logistique sont utilisés pour la discrimination (paragraphe 5). Enfin, nous présentons brièvement la macro DLOGISTIC (paragraphe 6), avant de conclure (paragraphe 7).

Pour illustrer les différentes notions développées, nous reprenons l'exemple proposé par ALBERT et HARRIS [1987], concernant des patients souffrant de quatre types de maladie du foie et sur lesquels quatre dosages d'enzymes ont été réalisés. Le quatrième dosage étant sans utilité pour la prédiction du type de maladie, il ne sera plus considéré par la suite. On se trouve donc en présence d'un jeu de données comportant quatre variables : la variable précisant la pathologie du patient, c'est-à-dire l'appartenance de l'individu à un groupe donné, et trois variables correspondant aux trois dosages retenus.

2. La régression logistique binaire

On considère n individus, dont n_1 présentent un caractère donné et dont les n_2 autres ne présentent pas ce caractère. La présence ou non du caractère est décrite par une variable y à deux modalités. Une telle variable est dite binaire. Par ailleurs, ces individus sont également décrits par p variables. L'objectif de la régression logistique est de modéliser la probabilité d'avoir le caractère donné, compte tenu des valeurs qui sont observées pour les p variables, dites variables explicatives.

Pour illustrer cette régression logistique binaire nous prenons l'exemple décrit dans l'introduction, en nous limitant aux 117 patients atteints par les deux dernières pathologies : 40 patients sont atteints de la pathologie numérotée 3 et 77 patients sont atteints de la pathologie numérotée 4. La variable à expliquer comporte donc 40 valeurs « 3 » et 77 valeurs « 4 » et on dispose de trois variables explicatives notées x_1, x_2 et x_3. Nous décidons en outre, de manière arbitraire, que la pathologie codée 4 correspond à la présence du caractère et que, par conséquent, la pathologie codée 3 correspond à l'absence du caractère.

On appelle « événement » le fait de présenter le caractère, c'est-à-dire, pour l'exemple, de présenter la pathologie codée 4.

1. Faculté universitaire des Sciences agronomiques de Gembloux (Belgique)

La probabilité de l'événement pour un individu j caractérisé par un vecteur d'observations $\boldsymbol{x}_j$ est notée $\pi(\boldsymbol{x}_j)$.

En régression logistique, on modélise, non pas $\pi(\boldsymbol{x}_j)$, mais une transformation de $\pi(\boldsymbol{x}_j)$. La transformation utilisée est la transformation logit, notée $g(\pi(\boldsymbol{x}_j))$:

$$g\left(\pi\left(\boldsymbol{x}_j\right)\right) = \ln\left(\frac{\pi\left(\boldsymbol{x}_j\right)}{1-\pi\left(\boldsymbol{x}_j\right)}\right).$$

Pour l'exemple considéré on a :

$$g\left(\pi\left(\mathbf{x}_j\right)\right) = \ln\left(\frac{P\left(y_j = 4 \mid \boldsymbol{x}_j\right)}{P\left(y_j = 3 \mid \boldsymbol{x}_j\right)}\right).$$

Le logit est donc le rapport de la probabilité que l'individu présente un caractère donné, considéré comme l'événement, sur la probabilité qu'il ne présente pas ce caractère, compte tenu de ses caractéristiques observées $\boldsymbol{x}_j$.

Les logits sont exprimés en fonction des caractéristiques des individus par le modèle de régression linéaire suivant :

$$g\left(\pi\left(\boldsymbol{x}_j\right)\right) = \alpha + \boldsymbol{\beta}\boldsymbol{x}_j + \varepsilon.$$

La constante α et le vecteur des paramètres $\boldsymbol{\beta}$ sont estimés, généralement par une méthode itérative, basée sur le maximum de vraisemblance. On peut ensuite estimer les probabilités $\hat{\pi}\left(\boldsymbol{x}_j\right)$ par la transformation inverse de la transformation logit :

$$\hat{\pi}\left(\mathbf{x}_j\right) = \frac{e^{g\left(\hat{\pi}\left(\mathbf{x}_j\right)\right)}}{1+e^{g\left(\hat{\pi}\left(\mathbf{x}_j\right)\right)}}.$$

Pour l'exemple considéré, on obtient le résultat suivant :

$$g\left(\hat{\pi}\left(x_j\right)\right) = 2{,}53774 + 0{,}01188x_1 - 0{,}02774x_2 - 0{,}04611x_3.$$

Ainsi, pour un patient caractérisé par les teneurs suivantes :

$$x_1 = 80 \quad x_2 = 110 \quad \text{et} \quad x_3 = 20,$$

on obtient :

$$g\left(\hat{\pi}\left(\mathbf{x}_j\right)\right) = -0{,}485,$$

et

$$\hat{\pi}\left(\mathbf{x}_j\right) = \frac{e^{-0{,}485}}{1+e^{-0{,}485}} = 0{,}381.$$

On peut donc conclure qu'un patient, atteint de la pathologie 3 ou 4 et pour lequel les teneurs pour les trois enzymes considérés seraient de 80, 110 et 20, aurait une probabilité de 0,38 d'être atteint de la pathologie 4 et une probabilité de 0,62 d'être atteint de la pathologie 3. Les probabilités peuvent évidemment être déterminées de façon similaire pour n'importe quel individu et notamment pour tous les individus de l'échantillon. Elles sont appelées probabilités *a posteriori* et correspondent à des probabilités conditionnelles, c'est-à-dire pour un vecteur $\boldsymbol{x}_j$ fixé, soit, pour l'exemple, pour trois valeurs fixées des teneurs en enzymes.

On notera cependant que, pour que les estimations des probabilités *a posteriori* soient satisfaisantes, il faut, d'une part, que le modèle de régression soit adéquat et, d'autre part, que la sélection des individus de l'échantillon ait été réalisée selon une procédure particulière.

Nous ne détaillons pas ici les différents critères qui peuvent être utilisés pour vérifier l'adéquation du modèle. Il s'agit d'un problème tout à fait comparable à celui qu'on rencontre en régression multiple classique. Pour des informations à ce sujet, nous renvoyons le lecteur aux ouvrages consacrés à la régression logistique, comme le livre de HOSMER et LEMESHOW [2000]. Une note est également disponible sur le site de l'Unité [DUYME et CLAUSTRIAUX, 2006].

Quant aux aspects liés à l'échantillonnage, nous y reviendrons au paragraphe 4.

3. La régression logistique polychotomique nominale

La régression logistique peut être étendue au cas où les individus sont répartis en plus de deux groupes : la variable *y* présente alors *g* modalités ($g > 2$), qui ne sont pas ordonnées. Ainsi, pour l'exemple décrit dans l'introduction, les patients peuvent présenter quatre pathologies différentes, notées 1, 2, 3 et 4.

En régression binaire, on a défini pour chaque individu un seul logit, qui est le rapport du logarithme de la probabilité de voir se réaliser l'événement sur la probabilité de ne pas voir se réaliser l'événement. Il s'agit donc du logarithme du rapport des probabilités relatives aux deux modalités, l'une étant considérée comme l'événement et l'autre comme la référence.

Pour la régression polychotomique, on définit $g - 1$ logits par individu, en exprimant chaque fois la probabilité relative à chacune des $g - 1$ modalités par rapport à une même modalité, choisie comme référence.

Ainsi, pour l'exemple, en prenant arbitrairement la pathologie 1 comme référence, on a :

$$g_{2/1}\left(\pi\left(\boldsymbol{x}_j\right)\right) = \ln\left(\frac{P\left(y=2 \mid \boldsymbol{x}_j\right)}{P\left(y=1 \mid \boldsymbol{x}_j\right)}\right),$$

$$g_{3/1}\left(\pi\left(\boldsymbol{x}_j\right)\right) = \ln\left(\frac{P\left(y=3 \mid \boldsymbol{x}_j\right)}{P\left(y=1 \mid \boldsymbol{x}_j\right)}\right)$$

et

$$g_{4/1}\left(\pi\left(\boldsymbol{x}_j\right)\right)=\ln\left(\frac{P\left(y=4\,\middle|\,\boldsymbol{x}_j\right)}{P\left(y=1\,\middle|\,\boldsymbol{x}_j\right)}\right).$$

Les différents logits sont alors exprimés en fonction de $\boldsymbol{x}_j$, par trois modèles linéaires :

$$g_{2/1}\left(\hat{\pi}\left(x_j\right)\right)=\hat{\alpha}_{2/1}+\hat{\boldsymbol{\beta}}_{2/1}\boldsymbol{x}_j,$$

$$g_{3/2}\left(\hat{\pi}\left(x_j\right)\right)=\hat{\alpha}_{3/1}+\hat{\boldsymbol{\beta}}_{3/1}\boldsymbol{x}_j,$$

et

$$g_{4/2}\left(\hat{\pi}\left(x_j\right)\right)=\hat{\alpha}_{4/1}+\hat{\boldsymbol{\beta}}_{4/1}\boldsymbol{x}_j,$$

et ces fonctions permettent de retrouver les probabilités *a posteriori* d'appartenance à chaque groupe :

$$P\left(y=1/\,\boldsymbol{x}_j\right)=\frac{1}{1+e^{g_{2/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{3/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{4/1}\left(\hat{\pi}\left(x_j\right)\right)}},$$

$$P\left(y=2/\,\boldsymbol{x}_j\right)=\frac{e^{g_{2/1}\left(\hat{\pi}\left(x_j\right)\right)}}{1+e^{g_{2/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{3/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{4/1}\left(\hat{\pi}\left(x_j\right)\right)}},$$

$$P\left(y=3/\,\boldsymbol{x}_j\right)=\frac{e^{g_{3/1}\left(\hat{\pi}\left(x_j\right)\right)}}{1+e^{g_{2/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{3/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{4/1}\left(\hat{\pi}\left(x_j\right)\right)}},$$

$$P\left(y=4/\,\boldsymbol{x}_j\right)=\frac{e^{g_{4/1}\left(\hat{\pi}\left(x_j\right)\right)}}{1+e^{g_{2/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{3/1}\left(\hat{\pi}\left(x_j\right)\right)}+e^{g_{4/1}\left(\hat{\pi}\left(x_j\right)\right)}}.$$

Les modèles obtenus sont les suivants :

$$g_{2/1}\left(\hat{\pi}\left(\boldsymbol{x}_j\right)\right)=3{,}06538-0{,}00244x_1-0{,}00940x_2-0{,}06192x_3,$$

$$g_{3/1}\left(\hat{\pi}\left(\boldsymbol{x}_j\right)\right)=-1{,}12496+0{,}03363x_1-0{,}03117x_2+0{,}23442x_3,$$

et

$$g_{4/1}\left(\hat{\pi}\left(\boldsymbol{x}_j\right)\right)=1{,}92702+0{,}05493x_1-0{,}07268x_2+0{,}17394x_3,$$

et, pour le patient considéré au paragraphe 2 et caractérisé par les dosages suivants :

$$x_1 = 80, \quad x_2 = 110 \quad \text{et} \quad x_3 = 20,$$

on a :

$$g_{2/1}\left(\hat{\pi}\left(x_j\right)\right) = 0{,}598, \quad g_{3/1}\left(\hat{\pi}\left(x_j\right)\right) = 2{,}825 \text{ et } g_{4/1}\left(\hat{\pi}\left(\boldsymbol{x}_j\right)\right) = 1{,}806,$$

et les probabilités *a posteriori* valent :

$$P\left(y = 1/\boldsymbol{x}_j\right) = 0{,}039, \quad P\left(y = 2/\boldsymbol{x}_j\right) = 0{,}071, \quad P\left(y = 3/\boldsymbol{x}_j\right) = 0{,}654, \quad P\left(y = 4/\boldsymbol{x}_j\right) = 0{,}236\,.$$

Comme pour la régression binaire, différents critères permettent de vérifier l'adéquation du modèle. Des informations à ce sujet sont données par HOSMER et LEMESHOW [2000], notamment.

De même, des informations sur la procédure d'échantillonnage supposée avoir été utilisée sont données au paragraphe suivant.

4. Procédures d'échantillonnage

Aux paragraphes 2 et 3, nous avons, à partir des fonctions logistiques, estimé les probabilités *a posteriori* d'appartenance à un groupe donné. Nous avons également signalé que ces estimations ne sont correctes que si la sélection des individus de l'échantillon a été réalisée selon une procédure adéquate.

En effet, le modèle régression logistique, binaire ou polychotomique, et les estimations qui en découlent, supposent que l'échantillonnage est aléatoire conditionnellement à $\boldsymbol{x}$, comme c'est également le cas en régression classique. Cela signifie que les vecteurs $\boldsymbol{x}_j$ peuvent être fixés arbitrairement et qu'un individu est choisi aléatoirement parmi tous les individus présentant le même vecteur $\boldsymbol{x}_j$ et qu'on observe le groupe auquel l'individu sélectionné se rattache.

Un tel schéma d'échantillonnage est parfois appelé schéma prospectif. Il se rencontre typiquement lorsque les variables explicatives correspondent à des niveaux ou à des variantes de facteurs étudiés dans un plan d'expérience.

Le modèle de régression est également applicable lorsqu'on choisit aléatoirement des individus dans une population et qu'on observe ensuite, sur chaque individu, le vecteur $\boldsymbol{x}_j$ et le groupe auquel appartient l'individu. Un tel schéma d'échantillonnage est appelé schéma de mélange, puisque l'échantillonnage se fait dans la population globale, constituée des sous-populations correspondant aux groupes.

Un troisième schéma d'échantillonnage que l'on rencontre en pratique est le schéma d'échantillonnage séparé dans les groupes. Contrairement aux deux schémas précédents, l'échantillonnage séparé dans les groupes fait intervenir l'appartenance aux groupes lors de la sélection : on sélectionne au hasard n_1 individus dans la sous-population des individus présentant la modalité 1, n_2 individus dans la sous-population des individus présentant la modalité 2, et ainsi de suite, s'il y a plus de deux modalités.

Ce schéma, appelé schéma rétrospectif, a comme conséquence que le praticien fixe la taille des sous-échantillons. Il en résulte que la proportion d'un groupe au sein de l'échantillon global ne reflète pas nécessairement la proportion d'un groupe dans la population, qui est appelée probabilité *a priori* d'appartenance au groupe. Dans une telle situation, les probabilités d'appartenance *a posteriori*, telles que calculées précédemment ne sont plus valables, mais devront être corrigées, en tenant compte des probabilités *a priori*, comme nous allons le voir au paragraphe suivant.

On peut montrer qu'en régression logistique, le fait de modifier la proportion d'individus d'une modalité par rapport aux autres n'a pas d'incidence sur les coefficients de régression des variables explicatives mais bien sur les ordonnées à l'origine. C'est donc uniquement les ordonnées à l'origine des modèles qu'il faudra modifier pour prendre en compte les proportions et les probabilités *a priori*. Il s'agit là d'une propriété spécifique à la transformation logit qui est utilisée dans la régression logistique.

5. Règle de classement des individus

Le principe du classement des individus dans les groupes est très simple : on classe un individu dans le groupe pour lequel la probabilité *a posteriori* est la plus grande.

Cela suppose donc qu'on puisse déterminer ces probabilités. Or, nous avons dit au paragraphe précédent que les probabilités *a posteriori* données par les programmes de régression sont influencées par les proportions d'observations réalisées pour les différents groupes. Plus précisément, ce sont les termes indépendants des fonctions logistiques qui doivent être corrigés si les effectifs des groupes dans les échantillons ne sont pas proportionnels aux probabilités *a priori*.

Si la fonction logistique du groupe *l* par rapport au groupe de référence *h*, obtenue par un programme de régression logistique, s'écrit :

$$g_{l/h}\left(\hat{\pi}\left(\boldsymbol{x}_j\right)\right) = \hat{\alpha}_{l/h} + \hat{\boldsymbol{\beta}}_{l/h}\boldsymbol{x}_j,$$

la fonction corrigée s'écrit :

$$g^*_{l/h}\left(\hat{\pi}\left(\boldsymbol{x}_j\right)\right) = \hat{\alpha}_{l/h} - \ln\left(\frac{n_l}{n_h}\right) - \ln\left(\frac{p_h}{p_l}\right) + \hat{\boldsymbol{\beta}}_{l/h}\boldsymbol{x}_j,$$

n_l et n_g étant les effectifs des échantillons des groupes *l* et *h* et p_l et p_h étant les probabilités *a priori* de ces groupes. On constate que cette correction s'annule si les effectifs des groupes sont proportionnels aux probabilités *a priori*.

La réalisation de la correction suppose donc que l'utilisateur dispose des probabilités *a priori*, ou du moins d'estimations de ces probabilités. Si l'échantillonnage des individus a été réalisé dans la population du mélange des groupes, les probabilités *a priori* peuvent être estimées par les proportions observées dans l'échantillon du mélange :

$$\hat{p}_l = \frac{n_l}{\sum_{i=1}^{g} n_i} \qquad (l = 1, ..., g),$$

Il y a, dans ce cas, automatiquement proportionnalité entre les effectifs des groupes et les probabilités *a priori* estimées et aucune correction des fonctions logistiques ne doit être réalisée. Cette situation est sans doute la plus simple car elle ne nécessite aucune information extérieure.

A titre d'illustration, nous reprenons l'exemple du paragraphe 3, en supposant cependant maintenant que les probabilités *a priori* ne sont plus proportionnelles aux effectifs des groupes mais égales à :

$$p_1 = 0{,}40, \quad p_2 = 0{,}20, \quad p_3 = 0{,}15 \quad \text{et} \quad p_4 = 0{,}25,$$

ces probabilités ayant, par exemple, été déterminées au cours d'une étude antérieure.

En désignant par $\hat{\alpha}_{l/1}$ et $\hat{\alpha}^*_{l/1}$ $(l = 2, ..., 4)$ les termes indépendants non corrigés et corrigés on a :

$$\hat{\alpha}^*_{2/1} = 3{,}06538 - \ln\left(\frac{44}{57}\right) - \ln\left(\frac{0{,}40}{0{,}20}\right) = 2{,}63110,$$

$$\alpha^*_{3/1} = -1{,}12496 - \ln\left(\frac{40}{57}\right) - \ln\left(\frac{0{,}40}{0{,}15}\right) = -1{,}75162,$$

et

$$\alpha^*_{4/1} = 1{,}92702 - \ln\left(\frac{77}{57}\right) - \ln\left(\frac{0{,}40}{0{,}25}\right) = 1{,}15626\,.$$

Pour le patient caractérisé par les trois teneurs suivantes :

$$x_1 = 80 \quad x_2 = 110 \quad \text{et} \quad x_3 = 20,$$

les trois logits valent :

$$g^*_{2/1}\left(\hat{\pi}(\boldsymbol{x}_j)\right) = 0{,}164, \quad g^*_{3/1}\left(\hat{\pi}(\boldsymbol{x}_j)\right) = 2{,}199 \quad \text{et} \quad g^*_{4/1}\left(\pi(\boldsymbol{x}_j)\right) = 1{,}035.$$

Il en résulte que les probabilités *a posteriori* pour cette personne sont égales à :

$$P(y = 1/\boldsymbol{x}_j) = 0{,}071, \quad P(y = 2/\boldsymbol{x}_j) = 0{,}841, \quad P(y = 3/\boldsymbol{x}_j) = 0{,}643 \quad \text{et} \quad P(y = 4/\boldsymbol{x}_j) = 0{,}201.$$

La règle de discrimination établie classe donc le patient dans le groupe 3, puisque c'est pour ce groupe que la probabilité d'appartenance est la plus élevée.

6. La macro DLOGISTIC

La macro DLOGISTIC réalise l'analyse discriminante logistique. Elle reclasse également les individus utilisés pour l'établissement de la règle de classement. Elle permet aussi de classer des individus non utilisés pour l'établissement de la règle.

Diverses options permettent de définir les probabilités *a priori*, de vérifier l'intérêt de la règle de classement établie, d'enregistrer les fonctions logistiques, les probabilités *a posteriori* et les groupes dans lesquels sont classés les individus.

Des informations complémentaires sont données dans la notice d'utilisation qui accompagne la macro.

La figure 1 donne les commandes et les résultats de leur exécution pour l'exemple traité au paragraphe 2.

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

%DLOGISTIC c32 c33 c34 c35;
    EQUATIONS c40;
    NOEFFIC;
    XVALID c36 c37 c38 c39 ;
    VPOSTERIORS c41 c42 .

Name c40 'g4/3'
Name c41 'APOSTERIORI3' c42 'APOSTERIORI4'
Print c40 c41 c42


RESULTATS
---------

 ANALYSE DISCRIMINANTE LOGISTIQUE

 RESULTATS POUR CALIBRAGE

Row  GROUPEOR   APRIORI  TABL.1  TABL.2  EFFECTI    TAUXER
  1         3  0.341880      25      15       40  0.375000
  2         4  0.658120       8      69       77  0.103896
  3                                      117  0.196581

 RESULTATS POUR CLASSEMENT INDIVIDUS ORIGINE INCONNUE

Row  TABL.1  TABL.2
  1       1       0



Row      g4/3  APOSTERIORI3  APOSTERIORI4
  1   2.53774      0.619033      0.380967
  2   0.01188
  3  -0.02774
  4  -0.04611
```

Figure 1. Analyse discriminante logistique dans le cas de deux groupes : commandes et résultats.

Pour les 117 patients, la variable à expliquer se trouve dans la colonne C32, et les variables explicatives dans les colonnes C33 à C35. Pour l'individu à classer, la variable définissant le groupe d'origine correspond à la colonne C36 et les variables explicatives correspondent aux colonnes C37 à C39. Dans la mesure où cet individu est d'origine inconnue, la valeur de la variable *y* a été fixée à une valeur différente des valeurs reprises dans la colonne C32 : les pathologies ont été numérotées 2 et 3 dans la colonne C32, mais 5 dans la colonne C36. L'équation de la fonction logistique est enregistrée dans la colonne C40 et les probabilités *a posteriori* du patient à classer dans les colonnes C41 et C42. Ces trois dernières colonnes ont, en outre, été imprimées.

Les probabilités *a priori* sont, par défaut, égales aux proportions des patients dans les deux groupes.

Le reclassement des 117 patients dans les deux groupes montre que 15 patients parmi les 40 patients du groupe 3 sont classés dans le groupe 4, soit un taux d'erreur pour le groupe 3 de 37,5 % ; 8 patients parmi les 77 patients du groupe 4 sont classés dans le groupe 3, soit un taux d'erreur de 10,4 %. Le taux d'erreur moyen pondéré par les probabilités *a priori* est donc égal à :

$$(0{,}342)\ (0{,}375) + (0{,}658)\ (0{,}104) = 0{,}197,$$

soit environ 20 %.

Quant au patient de pathologie indéterminée, la règle de classement qui a été définie le classe dans le groupe 3, la probabilité *a posteriori* pour ce groupe étant de 0,62.

On retrouve bien les valeurs qui ont été données au paragraphe 2.

La figure 2 concerne l'exemple du paragraphe 5. Pour les besoins de l'illustration, on a considéré en outre qu'on disposait de 40 individus supplémentaires, d'origine connue. Les données pour ces individus ont été générées artificiellement (colonnes C6 à C9) et sont utilisées pour montrer comment on pourrait valider la règle de décision construite sur les individus de calibrage (colonnes C2 à C5). Les probabilités *a priori* et les groupes auxquels ces individus sont affectés sont enregistrés dans la colonne C14 à C18.

La macro fournit deux tableaux de structure identique, donnant, notamment, pour chaque groupe d'origine (en lignes) les groupes dans lesquels les individus sont classés. Le premier tableau concerne les données de calibrage et le deuxième tableau concerne les données de validation. Ainsi, par exemple, pour les 57 patients de l'échantillon de calibrage qui présentent effectivement la pathologie 1, 53 sont reclassés dans ce groupe, trois sont classés dans le groupe 2 et 1 est classé dans le groupe 3. Pour cette pathologie, le taux d'erreur est de 7 %. Pour l'ensemble des groupes, le taux d'erreur moyen pondéré par les probabilités *a priori* est de 15,6 %.

Les fonctions logistiques ont été imprimées de même que les probabilités *a posteriori* pour les individus de validation. Toutefois, dans la figure 2, nous n'avons repris que les cinq premiers individus. L'examen des colonnes C6 à C9 montre que le premier patient est caractérisé par les trois teneurs suivantes :

$$x_1 = 80,\ x_2 = 110 \text{ et } x_3 = 20.$$

Il s'agit donc du patient pris en compte dans l'exemple du paragraphe 5 mais ici, nous avons considéré qu'on connaissait sa pathologie (groupe 3). On peut constater, ici aussi, qu'on retrouve bien les résultats donnés au paragraphe 5.

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

%DLOGISTIC c2 c3 c4 c5 ;
    PRIORS 0.4 0.20 0.15 0.25;
    EQUATIONS c11 c12 c13;
    XVALID c6 c7 c8 c9 ;
    NOEFFIC;
    VPOSTERIORS c14 c15 c16 c17;
    VGROUPE c18.

Name c11 'g2/1' c12 'g3/1' c13 'g4/1'
Name c14 'APOSTER1' c15 'APOSTER2' c16 'APOSTER3' c17 'APOSTER4'
NAME c18 'G_Affect'

print c11-c13
print c14-c18


RESULTATS
---------

 ANALYSE DISCRIMINANTE LOGISTIQUE

 RESULTATS POUR CALIBRAGE

Row  GROUPEOR  APRIORI  TABL.1  TABL.2  TABL.3  TABL.4  EFFECTI    TAUXER
  1         1     0.40      53       3       1       0       57  0.070175
  2         2     0.20       3      38       1       2       44  0.136364
  3         3     0.15       2       2      21      15       40  0.475000
  4         4     0.25       0       1       8      68       77  0.116883
  5                                                       218  0.155814

 RESULTATS POUR VALIDATION

Row  GROUPEOR  APRIORI  TABL.1  TABL.2  TABL.3  TABL.4  EFFECTI  TAUXER
  1         1     0.40       9       1       0       0       10    0.10
  2         2     0.20       1       8       0       1       10    0.20
  3         3     0.15       0       1       5       4       10    0.50
  4         4     0.25       0       0       1       9       10    0.10
  5                         10      10       6      14       40    0.18

Row      g2/1      g3/1      g4/1
  1   2.63110  -1.75162   1.15626
  2  -0.00244   0.03363   0.05493
  3  -0.00940  -0.03117  -0.07268
  4  -0.06192   0.23442   0.17394


Row  APOSTER1  APOSTER2  APOSTER3  APOSTER4  G_Affect
  1  0.071395  0.084104  0.643494  0.201007         3
  2  0.022120  0.099710  0.093066  0.785105         4
  3  0.030731  0.202591  0.058849  0.707829         4
  4  0.006322  0.004540  0.594506  0.394632         3
  5  0.949985  0.049794  0.000221  0.000000         1
```

Figure 2. Analyse discriminante logistique dans le cas de quatre groupes : commandes et résultats.

7. Conclusions

La macro proposée permet à l'utilisateur de réaliser l'analyse discriminante logistique avec la même facilité que l'analyse discriminante linéaire et quadratique. Le choix de la méthode ne sera donc plus guidé par les seules considérations purement techniques, mais pourra dépendre du respect ou non des conditions d'applications.

A ce sujet, rappelons que l'analyse discriminante linéaire est adaptée aux cas où les données des différents groupes proviennent de distributions multinormales, de même matrice de variances et covariances, que l'analyse quadratique est théoriquement adaptée aux cas de données multinormales de matrices de variances et de covariances inégales, mais que l'analyse discriminante logistique s'applique dans des conditions nettement moins restrictives. Dans une comparaison de ces trois méthodes, GLELE KAKAI et PALM [2007] ont montré la supériorité de l'analyse discriminante logistique sauf si l'hypothèse de normalité peut être acceptée et que l'hétéroscédasticité est nulle ou faible, auquel cas l'analyse discriminante linéaire est supérieure. Par contre, l'analyse quadratique ne s'est pas montrée supérieure à l'analyse logistique pour des populations normales de matrices de variances et covariances inégales.

Indépendamment de la discussion concernant la supériorité d'une méthode par rapport à une autre, il peut être utile, dans certains cas, de vérifier si la règle d'affectation qui est utilisée conduit à un taux de classement correct significativement supérieur à celui obtenu par une répartition des individus dans les groupes qui serait aléatoire, avec une proba bilité proportionnelle aux probabilités *a priori*, ou encore par une affectation systématique des individus au groupe pour lequel la probabilité *a priori* est maximale. Pour répondre à cette question, une macro spécifique a été mise au point et est incluse dans la macro DLOGISTIC. Les résultats fournis par cette macro n'ont cependant pas été repris dans les figures 1 et 2. Pour des informations complémentaires à ce sujet, nous renvoyons l'utilisateur à un autre document explicatif disponible sur le site de l'Unité [PALM 2008].

La macro DLOGISTIC utilise les commandes BLOGISTIC et NLOGISTIC de Minitab. Des problèmes de non convergence de l'algorithme d'estimation des paramètres peuvent se présenter, comme en régression logistique, en particulier lorsqu'il y a séparation complète des groupes, c'est-à-dire lorsqu'un groupe est parfaitement séparé d'un autre groupe dans l'espace des variables explicatives.

Enfin, on notera aussi que la macro DLOGISTIC n'accepte que des variables de type numérique. Si l'appartenance aux groupes est indiquée par une variable alphabétique, l'utilisateur devra, avant l'utilisation de DLOGISTIC, transformer cette variable alphabétique en variable numérique. De même, des variables explicatives représentant des variantes d'un facteur qualitatif ne peuvent pas être utilisées telles quelles, même si elles sont numériques. Une variable décrivant un facteur à k modalités devra, en effet, être remplacée par $k - 1$ variables indiquant la présence ou l'absence de la modalité en question sur les différents individus.

Bibliographie

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

DUYME F., CLAUSTRIAUX J. J. [2006]. La régression logistique binaire. *Notes stat. Inform.* (Gembloux) 2006/4, 24 p.

FISHER R. A. [1936]. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*. 7, 179-188.

GLELE KAKAI R., PALM R. [2007]. Data driven choice of the classification rule in discriminant analysis applied to Isoberlinia Stands. *West. Afr. Biophy. Biomath.*. 1, 21-38.

HOSMER D., LEMESHOW S. [2000]. *Applied logistic regression*. New York, Wiley, 392 p.

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

MCLACHLAN G. J. [1992]. *Discriminant analysis and statistical pattern recognition*. New York, Wiley, 544 p.

PALM R. [1999]. L'analyse discriminante décisionnelle : principes et applications. *Notes Stat. Inform.* (Gembloux). 99/4, 41 p.

PALM R. [2008]. Macro Minitab pour la vérification de l'intérêt d'une règle de décision en analyse discriminante <http.www.fsagx.ac.be/si/>.