NOTES DE STATISTIQUE ET D'INFORMATIQUE

2006/4

LA RÉGRESSION LOGISTIQUE BINAIRE

F. DUYME et J.J. CLAUSTRIAUX

Faculté universitaire des Sciences agronomiques
Unité de Statistique, Informatique et Mathématique appliquées
GEMBLOUX
(Belgique)

LA REGRESSION LOGISTIQUE BINAIRE

F. DUYME[(a)] et J.J CLAUSTRIAUX[(b)]

RÉSUMÉ

Cette note décrit et illustre les principes de base de la régression logistique binaire lorsque les variables explicatives sont continues.

SUMMARY

This note describes and illustrates the principles of binary logistic regression when predictors are continuous variables.

1. INTRODUCTION

La régression est une méthode incontournable en traitement des données en particulier dans une démarche de modélisation. Elle consiste à mettre en relation une variable à expliquer Y avec une ou plusieurs variables explicatives appelées prédicteurs.

La méthode la plus utilisée est la régression linéaire multiple qui s'applique lorsque la variable à expliquer est, comme les variables explicatives, continue.

Lorsque la variable *Y* n'est pas continue mais traduit l'appartenance à un groupe, il devient incorrect d'employer la régression classique à des fins de modélisation ou de prévision [HOSMER et LEMESHOW, 1989]. Un cas particulier est celui pour lequel la variable *Y* ne prend que deux valeurs qui signifient l'appartenance à une catégorie ou à un groupe d'individus. Dans ce cas, la méthode statistique adaptée est la régression logistique binaire. Il est également concevable d'utiliser l'analyse discriminante linéaire dont l'objet essentiel est le classement c'est-à-dire l'affectation d'individus à des groupes prédéfinis. Cependant, lorsque les conditions d'application de l'analyse discriminante ne sont pas réunies, il est préférable d'employer la régression logistique binaire [FAN et WANG, 1999; PRESS et WILSON, 1978;

TOMASSONE *et al.*, 1988] qui est sans aucun doute la forme la plus courante de régression logistique [RYAN, 2000].

Le développement de modèles pour données binaires a été favorisé par les besoins des biologistes [COLLETT, 1991]. En 1936, FISHER employait les termes d'analyse discriminante [VAN HOUWELINGEN et LE CESSIE, 1988]. Le domaine médical s'est ensuite intéressé à ces techniques pour des études épidémiologiques, des pronostics de maladies, etc. [COLLETT, 1991]. COX a joué un rôle fondamental dans la formulation du modèle de régression logistique, comme en témoigne, par exemple, son ouvrage écrit en 1970 [COX, 1970] ou celui écrit conjointement avec SNELL [COX et SNELL, 1989]. A l'heure actuelle, c'est dans les domaines médical et pharmaceutique que la régression logistique est la plus employée. Dans d'autres domaines, comme en économie, en agronomie, en sociologie, l'utilisation de la régression logistique est aussi présente; de nombreux exemples sont repris dans les ouvrages déjà mentionnés.

Cette note introduit la régression logistique binaire lorsque les prédicteurs sont des variables quantitatives. Ainsi, dans un premier paragraphe, nous aborderons les principes de la méthode, c'est-à-dire la présentation du modèle et son ajustement. Dans un second paragraphe, nous donnons une interprétation des coefficients du modèle, en particulier à travers l'*odds-ratio*. Nous exposons ensuite les principaux critères de mesure de l'adéquation du modèle au paragraphe 3. Nous évoquons au paragraphe 4 quelques précautions d'usage qu'il faut avoir à l'esprit avant de construire un modèle. Quelques informations complémentaires font l'objet d'un cinquième paragraphe.

A tout moment, nous illustrons nos propos à l'aide d'un exemple (annexe 1). Cet exemple a été exploité avec les logiciels Minitab (annexe 2) et SAS (annexe 3).

2. PRINCIPES DE LA MÉTHODE

2.1. Le modèle logistique

Le modèle de régression logistique fait partie d'une famille de modèles appelés modèles linéaires généralisés[1] décrits par exemple dans les ouvrages de MCCULLAGH et NELDER [1989] ou DOBSON [1990]. Un modèle linéaire généralisé est défini par une fonction de lien[2] notée g qui met en relation la composante aléatoire (vecteur Y) et la composante systématique (matrice X et vecteur des coefficients). La linéarité du modèle dépend du choix de g, donc de la transformation des valeurs de Y. Il existe

[1] En anglais : *generalized linear models*.
[2] En anglais : *link function*.

plusieurs fonctions de lien : *logit*, *probit*, *log-log* en particulier. D'après COLLETT [1991], la fonction *logit* est la plus employée essentiellement pour sa simplicité.

Afin d'illustrer nos propos, nous avons recueilli, pour 141 parcelles de lin, des informations sur la densité de levée et le rendement. Cette seconde variable n'est connue que sous forme de deux catégories : 0 pour les faibles rendements et 1 pour les forts rendements. La figure 1 représente la catégorie de rendement en fonction de la densité de levée (*dlevée*). Nous avons ensuite établi des classes de densité de levée et pour chacune d'elles déterminé la proportion de 1. La relation entre cette proportion et le centre de chaque classe est donnée à la figure 2.

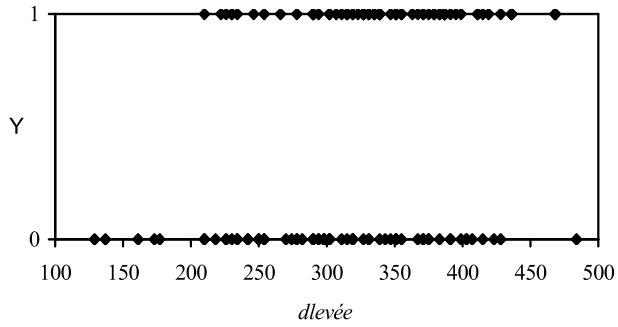

Figure 1. Catégories de rendement (*Y*) en fonction de la densité de levée (*dlevée*).

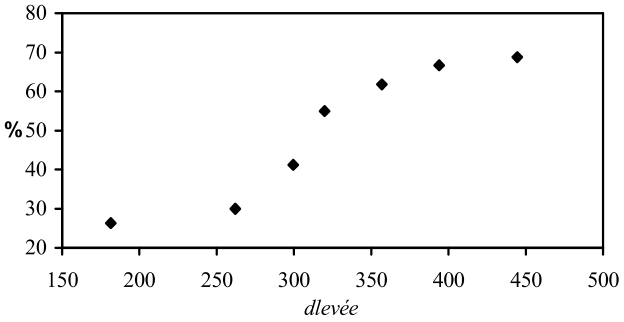

Figure 2. Fréquences relatives (%) de parcelles codées 1 par classe de densité de levée.

On constate que le graphique de la figure 2 a l'allure d'une distribution cumulée ou d'une fonction de répartition d'une variable aléatoire. Il n'est donc pas étonnant qu'une distribution connue puisse être utilisée pour modéliser la relation entre *Y* et la

densité de levée. La relation sigmoïdale, illustrée à la figure 2, peut être décrite par la fonction logistique qui se définit théoriquement comme :

$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}},$$

et est symétrique par rapport au point abscisse dont l'ordonnée vaut $f(x) = 0{,}5$.

Si $f(x)$ représente la probabilité pour un individu d'appartenir au groupe i des individus codés 1, probabilité notée π_i aussi appelée probabilité *a posteriori* ou probabilité de l'événement i, alors la forme du modèle logistique s'écrit:

$$\pi_i = \frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}},$$

où β_0 et β_1 sont les coefficients du modèle. La transformation de π_i utilisée s'appelle la transformation *logit*. Elle est donnée par la relation suivante :

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = g = \beta_0 + \beta_1 x_i$$

où *log* représente le logarithme népérien.

Lorsque p variables explicatives sont considérées sachant que j est l'indice des variables et i celui des individus, la fonction g s'écrit :

$$g = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

ou

$$g = \beta_0 + \sum_j \beta_j x_{ji} .$$

2.2. Ajustement du modèle

Ajuster un modèle de régression logistique revient à déterminer les coefficients $\beta_j \ (j = 0,\ldots,p)$ de la fonction g sur la base d'un échantillon de taille n.

Pour cela, nous utilisons la méthode du maximum de vraisemblance qui vise à fournir une estimation des paramètres qui maximise la probabilité d'obtenir les valeurs réellement observées sur l'échantillon [DAGNELIE, 1998; HOSMER et LEMESHOW, 1989]. Elle nécessite de définir une fonction de vraisemblance notée $L(\beta)$ et qui s'utilise sous la forme logarithmique :

$$log\, L(\beta) = \sum_i \left[y_i \, log\, \pi_i + (1 - y_i) log(1 - \pi_i) \right].$$

Cette équation dépend des probabilités π_i, inconnues, qui dépendent, à leur tour, des coefficients β_j inconnus. Afin de déterminer une estimation des β_j, on dérive l'expression $log\, L(\beta)$ en fonction des β_j. Cela nous amène, pour un prédicteur, à un système de deux équations dont la résolution ne peut se faire numériquement qu'au moyen d'une méthode itérative : il s'agit de la procédure des moindres carrés pondérés[3]. Il faut généralement plusieurs itérations avant d'obtenir une estimation par convergence des β_j.

Les 141 parcelles de lin ont été réparties aléatoirement en deux échantillons. Le premier, constitué de 71 parcelles, a permis d'ajuster un modèle de régression logistique dont l'équation est :

$$logit(\pi_i) = g = -4{,}412 + 0{,}01364 \; dlevee \; .$$

3. INTERPRÉTATION DES COEFFICIENTS

3.1. Généralités

Dans le modèle de régression apparaissent deux coefficients β_0 et β_1 encore appelés respectivement constante ou ordonnée à l'origine et pente ou coefficient de régression.

La pente de la droite représente le changement de *logit* pour un changement d'une unité de la variable *x* :

$$\hat{\beta}_1 = g(x+1) - g(x).$$

La valeur numérique de ce coefficient, comme celle de l'ordonnée à l'origine, n'ont pas d'interprétation directe. Seul, le signe du coefficient de régression permet de savoir si la probabilité de réussite est une fonction croissante ou non de la variable *x*.

[3] En anglais : *iterative weighted least squares, IWLS*

3.2. Rapport de chance

Le coefficient $\hat{\beta}_1$ sert à définir le rapport de chance ou *odds ratio* (*OR*). En effet, on peut montrer que :

$$\hat{\beta}_1 = g(x+1) - g(x) = logit(\hat{\pi}_{(x+1)}) - logit(\hat{\pi}_{(x)}) = log \frac{\hat{\pi}_{(x+1)}/1-\hat{\pi}_{(x+1)}}{\hat{\pi}_{(x)}/1-\hat{\pi}_{(x)}}.$$

C'est cette dernière quantité qui est souvent utilisée pour définir l'*odds ratio* :

$$\hat{OR} = exp(\hat{\beta}_1).$$

A titre d'illustration, pour un changement d'une unité de *dlevée*, on a :

$$\hat{OR} = exp(0{,}01364) = 1{,}014.$$

Cette valeur signifie que si on augmente la densité de levée d'une plante par m², alors la chance d'avoir un rendement plus élevé s'accroît de 1,4 %. Sur le plan agronomique et pour notre exemple, cela n'a pas de sens car la densité moyenne est proche de 300 plantes par m². Par contre, un accroissement de 50 plantes par m² procure une estimation de *OR* égale à 1,98 c'est-à-dire que les chances d'avoir un rendement élevé sont pratiquement doublées. C'est pourquoi, pour une variable continue, il est plus fréquent de calculer *OR* selon la formule :

$$\hat{OR} = exp(\delta\hat{\beta}_1)$$

où δ est la variation désirée. Davantage d'informations figurent dans l'ouvrage de HOSMER et LEMESHOW [1989].

4. PRINCIPAUX CRITÈRES DE QUALITÉ

4.1. Déviance et statistique de PEARSON

La fonction de vraisemblance résume l'information que les données apportent aux paramètres inconnus d'un modèle. Lorsque les paramètres inconnus sont égaux à leur estimation au sens du maximum de vraisemblance, la valeur de vraisemblance est une bonne mesure de la qualité d'ajustement des données par le modèle : c'est la vraisemblance maximale pour notre modèle, notée $L(\beta)$. Cette grandeur est cependant dépendante de la taille de l'échantillon. C'est pourquoi, elle est comparée à une

grandeur calculée sur les mêmes données mais pour un autre modèle appelé modèle saturé. Le modèle saturé comporte autant de paramètres qu'il y a d'observations dans l'échantillon. Dans ces conditions, les valeurs de $\hat{\pi}_i$ sont égales à celles de y_i. [COLLETT, 1991; HOSMER et LEMESHOW, 1989]. La vraisemblance du modèle saturé est notée *L(s)*.

Les deux quantités définissent la déviance[4] *D*:

$$D = -2\,log\left[\frac{L(\beta)}{L(s)}\right];$$

elles se déterminent comme suit :

$$D = -2\sum_i \left[y_i \; log\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\, log\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right].$$

Appliquée à notre exemple, cette formule conduit à une valeur de déviance égale à :

$$D = -2(-42{,}8) = 85{,}6.$$

La quantité -42,8 est souvent appelée *log* de vraisemblance[5]; la quantité 85,6 est désignée par -2*LogL*.

La déviance joue le même rôle que la somme des carrés des écarts résiduelle en régression classique. Sa valeur est d'autant plus petite que le modèle s'ajuste mieux aux données. Cependant, la forme donnée ci-dessus n'utilise que les valeurs de $\hat{\pi}_i$ et pas simultanément les valeurs de $\hat{\pi}_i$ et y_i. C'est pourquoi, la déviance n'est pas une bonne mesure d'adéquation du modèle.

Afin de tenir compte des valeurs de $\hat{\pi}_i$ et y_i, on utilise la statistique de PEARSON définie comme suit :

$$X^2 = \sum_i \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}.$$

[4] En anglais : *deviance*
[5] En anglais : *log likelihood*

La quantité X^2 est distribuée selon une variable χ^2 de PEARSON dont l'espérance mathématique ou nombre de degrés de liberté est *n-p-1*. Elle présente l'inconvénient de dépendre directement de la taille de l'échantillon et du nombre de classes. En pratique, plus cette quantité est petite et plus le modèle est adéquat.

Pour notre exemple, X^2 vaut 69,15 avec 69 degrés de liberté. Cependant, si on élimine automatiquement les répétitions des individus ayant les mêmes valeurs de *dlevée*, la valeur X^2 vaut 42,55 avec 39 degrés de liberté (logiciel Minitab).

Signalons encore que la racine carrée de chacune des quantité reprise dans la somme de la formule ci-dessus est appelée *résidu de PEARSON*.

4.2. Test de HOSMER et LEMESHOW

Le test de HOSMER et LEMESHOW permet d'éviter, au moins en partie, le problème lié à la taille (n) évoqué avec la statistique X^2.

HOSMER et LEMESHOW ont eu l'idée de regrouper les individus de manière à calculer une quantité proche de celle de PEARSON. Il existe plusieurs stratégies de regroupement. Celle basée sur la régularité des intervalles de probabilité, c'est-à-dire sur une répartition des individus en classes de probabilités de taille quasiment identique est plus judicieux, étant donné qu'elle est plus proche de la distribution théorique sous-jacente [HOSMER et LEMESHOW, 1989].

Les individus sont répartis en q classes (k=1, ...,q) et en deux groupes selon la valeur de Y (tableau 1). Les q classes sont définies de manière à obtenir des valeurs de n_k ou taille de la classe k identiques ou presque identiques. Pour le groupe 1, c'est-à-dire pour les individus codés 1, et par classes, on détermine les fréquences observées o_k et les fréquences attendues e_k par la somme des probabilités *a posteriori* pour tous ces individus. Dans les cellules du groupe 2, on pourrait reprendre des informations analogues qui ne sont pas utiles à considérer dans le cas de la régression logistique binaire puisque la différence entre n_k et o_k correspond au nombre d'individus codés zéro.

La statistique χ^2_{HL} de HOSMER et LEMESHOW se définit par :

$$\chi^2_{HL} = \sum_k \frac{(o_k - e_k)^2}{e_k(1 - e_k/n_k)},$$

Cette statistique est approximativement distribuée selon une variable χ^2 à q-2 degrés de liberté. Le nombre de degrés de liberté a été déterminé par l'intermédiaire de nombreuses simulations.

Tableau 1. Présentation schématique du regroupement des individus dans le cadre du test de HOSMER et LEMESHOW.

		Classes					
		1	2	k	….	q	
Groupes	1	o_1 e_1	o_2 e_2	o_k e_k	....	o_q e_q	
	2						
Totaux		n_1	n_2	n_k	....	n_q	n

Le nombre de classes de regroupement n'est pas fixe. Il est par exemple fixé par défaut à 10 dans le logiciel Minitab si le nombre d'individus est suffisant. On parle alors des "déciles de risque"[6], surtout en médecine [HOSMER et LEMESHOW, 1989]. Si n est petit, un minimum de trois classes est obligatoirement respecté afin de réaliser le test. HOSMER et LEMESHOW préconisent de ne pas descendre en dessous de six classes.

Pour notre exemple, à partir d'une répartition en 10 classes, nous obtenons une valeur de χ^2_{HL} égale à 10,23, avec huit degrés de liberté et une P-valeur de 0,25. Plusieurs valeurs de e_k sont cependant inférieures à 2. Une répartition en sept classes conduit à une statistique égale à 4,32 et une P-valeur de 0,50. Puisque les écarts entre les valeurs de o et de e (donc de y_i et de π_i) ne sont pas significatives, nous concluons à la bonne adéquation de notre modèle.

4.3. Rapport de vraisemblance

Une autre statistique couramment calculée par les logiciels est le rapport de vraisemblance[7] noté G :

$$G = -2\,log\left[\frac{L(0)}{L(\beta)}\right]$$

[6] En anglais : *deciles of risk*
[7] En anglais : *likelihood ratio*

où *L(0)* est la vraisemblance du modèle sans variable explicative. Cette quantité est également la différence entre la déviance du modèle sans variable et celle du modèle avec la variable. *G* est distribué selon un χ^2 à un degré de liberté. Lorsque *p* variables sont utilisées la distribution de référence est un χ^2 à p degrés de liberté. Plus la valeur de G est grande, plus la ou les variables utilisées dans le modèle sont intéressantes pour expliquer par exemple la catégorie des rendements élevés.

Dans le cas de notre exemple, *G* vaut 12,80 et la P-valeur 0,0 %. La variable (*dlevée*) est donc tout à fait intéressante.

4.4. Test de WALD

En régression classique, l'hypothèse de nullité de chaque coefficient est testée au moyen de la statistique de STUDENT. Le test équivalent en régression logistique s'appelle le test de WALD. Cette statistique, notée *W*, est obtenue en rapportant l'estimation $\hat{\beta}$ du maximum de vraisemblance à une estimation de son erreur standard (*se*) :

$$W = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})}$$

Pour *n* grand, *W* est approximativement distribué selon une loi normale réduite. Cette statistique est notée *z* dans Minitab, *WALD chi-square* dans SAS, ce qui correspond à la quantité z^2 distribuée de ce fait selon une variable χ^2 à un degré de liberté.

Toujours pour notre exemple, nous obtenons, pour *dlevée*, une valeur de *z* égale à 3,18 (10,04 dans SAS) et une P-valeur égale à 0,1 %. Nous concluons à la signification de cette variable.

Notons pour terminer qu'il existe également un test sur la constante. Par ailleurs, le test de WALD multivarié est assimilable, lorsque la taille de l'échantillon est suffisante, à un test sur le rapport de vraisemblance. Ce test est par exemple fourni par le logiciel SAS.

4.5. Autres critères

Il existe d'autres critères de mesure de l'adéquation d'un modèle. Ces critères ne sont pas tous déterminés par les logiciels, ni même détaillés dans un seul ouvrage.

Une statistique courante en régression linéaire classique s'appelle le coefficient de détermination R^2. Une variante, le $\overline{R}^2$ ou **coefficient de détermination ajusté**,

permet de tenir compte de la structure du modèle. RYAN [1997] montre que ces grandeurs ne peuvent pas être utilisées en l'état en régression logistique. Pour des modèles de régression logistique, la forme la plus simple du coefficient de détermination est donnée par la relation [MENARD, 2000] :

$$R_r^2 = \frac{logL(0) - logL(\beta)}{logL(\beta)}.$$

Une autre forme connue sous le nom de **coefficient de détermination généralisé** est définie par COX et SNELL [1989] :

$$R_L^2 = 1 - \left[\frac{L(0)}{L(\beta)}\right]^{2/n}.$$

Seule, cette seconde forme est calculée dans SAS.

Le **taux de bon classement** ou *CCR*[8] est une statistique courante pour déterminer le pourcentage d'individus correctement classés en appliquant le modèle de régression. Il permet de calculer le taux de mauvais classement de ce modèle. La probabilité *a posteriori* est calculée pour chaque individu. Elle est ensuite transformée en code binaire selon qu'elle dépasse ou non un seuil de partage[9] souvent fixé sans règle bien définie. On dresse ensuite un tableau 2×2 (tableau 2).

Le taux global de bon classement est le rapport $(n_{11} + n_{22})/n$. On définit également le taux de bon classement du premier groupe ou sensibilité[10] par n_{11}/n_1 et le taux de bon classement dans le second groupe ou spécificité[11] par n_{22}/n_2 [SAS, 1994b]. Sensibilité et spécificité trouvent leur origine en médecine lors du diagnostic de maladie. Un test est dit sensible si toutes les personnes réellement malades (code 1) sont classées dans le groupe des malades. Il est spécifique si toutes les personnes saines se trouvent dans le deuxième groupe [KOTZ et JOHNSON, 1988; SAS, 1995]. La proportion de faux positifs se définit par n_{21}/n'_1 et celle de faux négatifs par n_{12}/n'_2.

[8] En anglais : *correct classification rate, CCR*
[9] En anglais *cutpoint*
[10] En anglais : *sensitivity*
[11] En anglais : *specificity*

Tableau 2. Répartition des individus en fonction du groupe observé et du groupe prédit; n11, n12, n21, n22 : effectifs partiels.

Groupes		Prédits		Totaux
		1	0	
Observés	1	n_{11}	n_{12}	n_1
	0	n_{21}	n_{22}	n_2
Totaux		n'_1	n'_2	n

Les **mesures d'association**, bien qu'intégrées dans certains logiciels comme SAS et Minitab, sont très peu mentionnées dans la littérature relative à la régression logistique. Ces mesures sont en réalité des coefficients de corrélation des rangs et constituent de ce fait des statistiques non paramétriques. Quatre statistiques, *D* de SOMER, gamma de GOODMAN et KRUSKAL, Tau-a de KENDALL et *c* (ou surface sous la courbe *ROC*) sont principalement utilisées. Elles sont basées sur des paires concordantes, c'est-à-dire sur le nombre de fois où une probabilité $\hat{\pi}_i$ (d'un individu codé 1) est supérieure à celle d'un individu codé 0. Davantage d'informations figurent dans SAS [1995].

Nous terminons ce paragraphe par un dernier critère, le **score**. Ce critère est totalement absent des sorties de Minitab alors que dans SAS nous le trouvons à de nombreuses reprises. Le test du score est tout à fait similaire à celui du rapport de vraisemblance ou du test de WALD multivarié. Il ne faut cependant pas le confondre avec le score résiduel calculé sur les variables non incluses dans un modèle, lors d'une étape de sélection de variable. Pour une variable explicative, la forme de la statistique est donnée par HOSMER et LEMESHOW [1989]. Dans le cas de plusieurs prédicteurs, la forme explicite du test est rappelée, tout comme celle des deux autres critères, par DUYME [2001].

Pour notre exemple, nous obtenons les résultats suivants :

- $R_r^2 = 0{,}15$
- $R_L^2 = 0{,}16$
- CCR = 66 %
- D = 0,45
- gamma = 0,45
- tau-a = 0,23
- c = 0,73
- score = 11,8.

4.6. Validation externe

Après avoir ajusté un modèle, il est courant de le valider sur d'autres données, à condition d'en disposer. Sur un échantillon de faible taille, il est intéressant d'employer la validation croisée, le *jacknife* ou le *boostrap* afin de déterminer une estimation sans biais du taux de mauvais classement ou *TMC* [CELEUX, 1994; TOMASSONE *et al*., 1988].

Lorsque de nombreux individus sont disponibles, on crée un échantillon de validation de manière aléatoire (cet échantillon représente en général 25 à 50 % de l'ensemble des individus disponibles). Ainsi, à côté du *TMC*, la statistique de HOSMER et LEMESHOW ainsi que les critères d'association peuvent être déterminés sans biais.

A titre d'exemple, sur les 70 autres parcelles de lin, à partir du modèle obtenu sur l'autre échantillon, nous avons calculé les probabilités *a posteriori*. Ces probabilités sont ensuite codées en 1 ou 0 selon un seuil de partage fixé à 0,5. Nous obtenons, en croisant ces codes aux catégories initiales de rendement, un taux global de mauvais classement de 43 %, ce qui est élevé (45 % pour le groupe des 0 et 32 % pour le groupe des 1). Ainsi, notre modèle qui, en ajustement présentait de bons résultats, aurait, d'après ce critère, une qualité de prédiction plutôt médiocre.

5. PRÉCAUTIONS D'USAGE

L'emploi de la régression logistique nécessite le respect d'un certain nombre de précautions.

1° Notons tout d'abord qu'il n'est pas toujours possible de déterminer précisément des estimations du maximum de vraisemblance. Ainsi, nous avons, à titre d'illustration, utilisé une partie des données sur les iris de FISHER [SAS, 1994a] afin d'établir un modèle de régression. Ce fut sans succès, l'algorithme ne venant pas à converger. Nous sommes en réalité en situation de séparation complète[12] des deux espèces dans l'espace de la variable explicative entraînant une **non-convergence**. Cette situation est représentée à la figure 3 pour les 50 individus de chaque espèce. Les espèces forment deux groupes distincts, ce qui explique que l'algorithme *IWLS* ne parvienne pas à positionner facilement un modèle logistique : il ne peut estimer précisément les coefficients du modèle. Les cas de non-convergence sont assez facilement détectés : message sur la sortie d'ordinateur, estimation des coefficients β_j et/ou des écarts-types non fiable (généralement ces derniers ont une très grande valeur). La figure 4 fournit un exemple de sortie Minitab pour ce cas de non-convergence.

[12] En anglais : *complete separation*

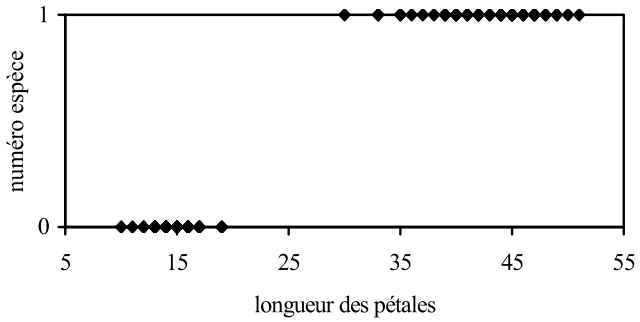

Figure 3. Espèces d'iris en fonction de la longueur des pétales : illustration d'un cas de séparation complète.

```
* REMARQUE
      * L'algorithme n'a pas convergé après 20 itérations.
      * La convergence n'a pas été atteinte pour le
      * critère d'estimations de paramètres.
      * Les résultats ne sont peut-être pas fiables.
      *    Essayer    d'augmenter    le    nombre    maximum
d'itérations.


Prédicteur         Coef    Ecart-type         Z      P
Constante           -43          4587     -0.01  0.993
V                    13          1294      0.01  0.992
```

Figure 4. Exemple de non-convergence avec le logiciel Minitab.

2° La **taille *n* de l'échantillon** est importante à prendre en considération. En effet, la distribution de la statistique de WALD n'est qu'asymptotique, ce qui signifie qu'il est nécessaire d'avoir une valeur de *n* élevée, idéalement, une centaine d'individus. En deçà de 50, il faut être prudent en régression multiple, en particulier sur la précision de l'estimation des β_j.

De même, des précautions doivent être prises si le nombre de variables explicatives est voisin ou supérieur à *n*. Néanmoins, nous pensons que cette situation est exceptionnelle si *n* est supérieur à 50.

3° Une attention doit également être portée sur la **proportion d'individus codés 1** [DUYME, 2001]. Si elle est trop faible (ou trop forte), il y a instabilité de la valeur des coefficients, manque de précision des estimations et à nouveau des cas de

non-convergence peuvent survenir. Une valeur minimale de 5 % doit impérativement être respectée. Il existe à ce sujet un critère appelé *EPV*[13] ou nombre de 1 par prédicteur. D'après les travaux de PEDUZZI *et al*. [1996], un *EPV* minimum de 10 est souhaité.

Au tableau 3, nous donnons les valeurs *n/p* ou nombre d'individus par variable explicative en fonction de la proportion de 1 $\left(n_1/n\right)$ et du nombre d'*EPV* $\left(n_1/p\right)$.Ces valeurs permettent de calculer *n* ou *p*. Ainsi, si on souhaite un rapport n_1/n égal à 10 % et un *EPV* de 10, le tableau indique que le rapport *n/p* doit être de 100. Dès lors, si on dispose de deux prédicteurs, par exemple, l'effectif *n* doit être de 200 et, par contre, de 1.000 pour dix variables explicatives.

Tableau 3. Nombre d'individus minimum par prédicteur (n/p) en fonction de n_1/n et de EPV (en gris : zone non recommandée).

		n_1/n				
		5	10	15	20	50
	1	20	10	7	5	2
	5	100	50	33	25	10
	10	200	100	67	50	20
EPV	14	280	140	93	70	28
	18	360	180	120	90	36
	22	440	220	147	110	44
	26	520	260	173	130	52

4° Parmi les autres précautions, il est important de mentionner la recherche des **données extrêmes**. Il existe pour cela des critères permettant un diagnostic fiable comme par exemple les résidus de PEARSON. Ces critères sont décrits par HOSMER et LEMESHOW [1989] ou COLLETT [1991], et ils sont calculés dans certains logiciels comme SAS [1994b]. Au tableau 4, nous donnons la valeur de quelques résidus de PEARSON déterminés à partir du modèle *dlevée* sur l'échantillon de 71 parcelles. Ces valeurs illustrent simplement le fait que les données extrêmes se situent non pas aux extrémités des valeurs de *dlevée* mais, d'après la figure 1, au niveau de la zone commune entre les deux groupes

5° Un dernier point important consiste à repérer la présence éventuelle de **colinéarité**. Alors qu'en régression classique, un diagnostic sur la matrice *X* suffit, en régression logistique il conviendrait de travailler sur la matrice information, obtenue à partir des matrices *X* et *Y*. Un résumé est fourni par DUYME [2001]. Lorsque la

[13] En anglais : *events per variable*

matrice est très colinéaire (on parle de *ill-conditioning*), il existe des méthodes d'ajustement alternatives à celle du maximum de vraisemblance. On peut à ce titre consulter BARKER et BROWN [2001].

Tableau 4. Résidus de PEARSON pour quelques couples (*dlevée*, y_i): exemple du lin (en gras : valeurs importantes).

y_i	*dlevée*	résidus
0	242	-0,84
0	90	-0,20
0	403	**-2,52**
1	407	0,57
1	230	**2,82**

6. COMPLÉMENTS D'INFORMATION

Tout comme en régression classique, le problème de la sélection de variables existe en régression logistique. Les mêmes méthodes de sélection sont employées. Elles sont décrites et illustrées dans SAS [1994b, 1995].

La régression logistique est également utilisée lorsque les données sont groupées. Ainsi, au lieu de se baser sur des individus ayant le code 0 ou 1, on a des proportions d'individus ayant le code 1 par valeur de la variable explicative. Cette variable est alors davantage catégorielle que quantitative et on parle de données binomiales [AGRESTI, 1990; COLLETT, 1991; LINDSEY, 1995].

Enfin, lorsque la variable dépendante possède plus de 2 réponses, on a affaire à la régression logistique polychotomique (ordinale ou nominale). Des informations intéressantes à ce sujet figurent dans l'ouvrage de HOSMER et LEMESHOW [1989]. Notons pour terminer que certains cas de régression polychotomique peuvent se simplifier en régression binaire, en tenant bien entendu compte de l'objectif de travail sous-jacent.

RÉFÉRENCES BIBLIOGRAPHIQUES

AGRESTI A. [1990]. *Categorical data analysis*. New York, Wiley, 558 p.

BARKER L., BROWN C. [2001]. Logistic regression when binary predictor variables are highly correlated. *Stat. Med*. **20**, 1431-1442.

CELEUX G. [1994]. Introduction générale. In: CELEUX et NAKACHE. *Analyse discriminante sur variables qualitatives.* Paris, *Polytechnica*, 1-17.

COLLETT D. [1991]. *Modelling binary data*. London, Chapman & Hall, 369 p.

COX D.R. [1970]. *The analysis of binary data*. London, Methuen.

COX D.R., SNELL E.J. [1989]. *Analysis of binary data*. London, Chapman & Hall, 236 p.

DAGNELIE P. [1998]. *Statistique Théorique et Appliquée*, vol. 1. Paris, De Boeck & Larcier, 508 p.

DOBSON A.J. [1990]. *An introduction to generalized linear models*. London, Chapman & Hall, 176 p.

DUYME F. [2001]. *Qualité des modèles de régression logistique binaire: effet de la proportion d'individus par catégorie et du mode d'utilisation des données* (thèse de doctorat). Gembloux, Faculté Universitaire des Sciences Agronomiques (Belgique); Paris-Grignon, Institut National Agronomique (France), 181 p.

FAN X., WANG L. [1999]. Comparing linear discriminant function with logistic regression for the two-group classification problem. *J. Experim. Edu*. **67**, 265-286.

HOSMER D.W., LEMESHOW S. [1989]. *Applied logistic regression*. New York, Wiley, 307 p.

KOTZ S., JOHNSON N.L. [1988]. *Encyclopedia of statistical sciences*, vol. 8. New York, Wiley, 870 p.

LINDSEY J.K. [1995]. *Modelling frequency and count data*. New York, Oxford University Press, 291 p.

McCULLAGH P., NELDER J.A. [1989]. *Generalized linear models.* London, Chapman & Hall, 511 p.

MENARD S. [2000]. Coefficients of determination for multiple logistic regression analysis. *Amer. Stat*. **54**, 17-24.

PEDUZZI P., CONCATO J., KEMPER E., HOLFORD T.R., FEINSTEIN A.R. [1996]. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin*. *Epidemiol*. **49**, 1373-1379.

PRESS S.J., WILSON S. [1978]. Choosing between logistic regression and discriminant analysis. *J. Amer. Stat. Assoc.* **73**, 699-705.

RYAN T.P. [1997]. *Modern regression methods*. New York, Wiley, 515 p.

RYAN T.P. [2000]. Some issues in logistic regression. *Comm. Stat.-Theo. Meth*. **29**, 2019-2032.

SAS Institute [1994a]. SAS/STAT- User's guide (proc DISCRIM), version 6, vol 1.

SAS Institute [1994b]. SAS/STAT- User's guide (proc LOGISTIC), version 6, vol 2.

SAS Institute [1995]. *Logistic regression: examples using the SAS system version 6*. Cary, SAS Institute Inc., 163 p.

TOMASSONE R., DANZART M., DAUDIN J.J., MASSON J.P. [1988]. *Discrimination et classement*. Paris, Masson, 172 p.

VAN HOUWELINGEN J.C., LE CESSIE S. [1988]. Logistic regression, a review. *Stat. Neerl*. **42**, 215-252.

ANNEXES

Annexe 1. Données relatives à la culture du lin; échantillon d'apprentissage (71 parcelles).

dlevée	*Y*	*dlevée*	*Y*
403	0	407	1
278	0	371	1
173	0	351	1
315	0	383	1
294	0	387	1
161	0	339	1
290	0	375	1
290	0	290	1
298	0	335	1
210	0	331	1
254	0	379	1
371	0	339	1
391	0	319	1
218	0	323	1
242	0	230	1
298	0	428	1
226	0	230	1
351	0	327	1
375	0	290	1
347	0	411	1
383	0	311	1
331	0	411	1
250	0	307	1
375	0	234	1
234	0	387	1
298	0	419	1
327	0	387	1
403	0	387	1
210	0	315	1
371	0	355	1
242	0	327	1
331	0	468	1
254	0	327	1
391	0	436	1
129	0	371	1
		399	1

dlevée = densité de levée
Y = catégorie de rendement (0: faible rendement; 1: fort rendement)

Annexe 2. Commandes et sortie Minitab.

```
MTB > BLogistic Y = dlevee;
SUBC>   Logit;
SUBC>   Hosmer 7;
SUBC>   Brief 2.

Fonction de liaison :  Logit

Informations de réponse

Variable  Valeur       Dénombrement
   Y      1                    36  (Evénement)
          0                    35
          Total                71

Tableau de régression logistique
                                                    Ratio de  IC à 95%
Prédicteur    Coef  Er-T coef        Z     P  probab. Infér Supér
Constante   -4,412      1,434    -3,08 0,002
dlevee       0,013636   0,004294    3,18 0,001  1,01   1,01  1,02

Log de vraisemblance = -42,806
Test de toutes les pentes sont zéro : G = 12,802; DL = 1; P = 0,000

Tests d'adéquation de l'ajustement

Méthode                 Khi deux    DL      P
Pearson                   42,554    39    0,321
Deviance                  51,294    39    0,090
Hosmer-Lemeshow            4,315     5    0,505

Tableau des effectifs observés et de leur espérance mathématique :
(voir le test de Hosmer-Lemeshow pour la statistique du Khi deux de
Pearson)

                          Groupe
Valeur     1     2     3     4     5     6     7   Total
1
  Obs      3     2     5     7     5     7     7      36
  Esp      1,9   3,2   4,5   5,3   7,0   7,0   7,1
0
  Obs      8     8     5     3     6     3     2      35
  Esp      9,1   6,8   5,5   4,7   4,0   3,0   1,9

  Total   11    10    10    10    11    10     9      71

Mesures d'association :
(entre la variable de réponse et les prévisions de probabilité)

Paires          Nombre Pourcentage  Mesures récapitulatives
Concordant         901       71,5%  D de Somers                0,45
Discordant         340       27,0%  Gamma de Goodman-Kruskal   0,45
Ex aequo            19        1,5%  Tau-a de Kendall           0,23
Total             1260      100,0%
```

Tableau des effectifs observés et de leur espérance mathématique:
(voir le test de Hosmer-Lemeshow pour la statistique du Khi deux de Pearson)

	Groupe							
Valeur	1	2	3	4	5	6	7	Total
1								
Obs	1	2	4	5	7	7	10	36
Esp	0.8	1.8	4.2	5.7	6.8	7.5	9.2	
0								
Obs	9	8	6	5	3	3	1	35
Esp	9.2	8.2	5.8	4.3	3.2	2.5	1.8	
Total	10	10	10	10	10	10	11	71

Mesures d'association :
(entre la variable de réponse et les prédictions de probabilité)

Paires	Nombre	Pourcentage	Mesures récapitulatives	
Concordant	1023	81.2%	D de Somers	0.63
Discordant	233	18.5%	Gamma de Goodman-Kruskal	0.63
Ex-aequo	4	0.3%	Tau-a de Kendall	0.32
Total	1260	100.0%		

Annexe 3. Commandes et sortie SAS.

```
proc logistic descending;
model Y=dlevee etract hflo idefmat/
lackfit rsq ctable;
run;
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.LIN1VAR
Response Variable	Y
Number of Response Levels	2
Number of Observations	71
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Y	Total Frequency
1	1	36
2	0	35

Probability modeled is Y=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	100.413	89.611
SC	102.675	94.136
-2 Log L	98.413	85.611

R-Square	0.1650	Max-rescaled R-Square	0.2200

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.8017	1	0.0003
Score	11.7989	1	0.0006
Wald	10.0839	1	0.0015

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.4121	1.4343	9.4625	0.0021
dlevee	1	0.0136	0.00429	10.0839	0.0015

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dlevee	1.014	1.005	1.022

Association of Predicted Probabilities and Observed Responses

Percent Concordant	71.5	Somers' D	0.445
Percent Discordant	27.0	Gamma	0.452
Percent Tied	1.5	Tau-a	0.226
Pairs	1260	c	0.723

Partition for the Hosmer and Lemeshow Test

Group	Total	Y = 1 Observed	Y = 1 Expected	Y = 0 Observed	Y = 0 Expected
1	7	0	1.03	7	5.97
2	7	3	1.66	4	5.34
3	7	2	2.46	5	4.54
4	8	3	3.48	5	4.52
5	9	6	4.61	3	4.39
6	7	5	4.01	2	2.99
7	7	3	4.63	4	2.37
8	7	6	4.88	1	2.12
9	8	4	5.95	4	2.05
10	4	4	3.29	0	0.71

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
10.2291	8	0.2493

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensitivity	Specificity	False POS	False NEG
0.060	36	0	35	0	50.7	100.0	0.0	49.3	.
0.080	36	1	34	0	52.1	100.0	2.9	48.6	0.0
0.100	36	1	34	0	52.1	100.0	2.9	48.6	0.0
0.120	36	2	33	0	53.5	100.0	5.7	47.8	0.0
0.140	36	3	32	0	54.9	100.0	8.6	47.1	0.0
0.160	36	3	32	0	54.9	100.0	8.6	47.1	0.0
0.180	36	3	32	0	54.9	100.0	8.6	47.1	0.0
0.200	33	5	30	3	53.5	91.7	14.3	47.6	37.5
0.220	33	7	28	3	56.3	91.7	20.0	45.9	30.0
0.240	33	8	27	3	57.7	91.7	22.9	45.0	27.3
0.260	33	10	25	3	60.6	91.7	28.6	43.1	23.1
0.280	33	11	24	3	62.0	91.7	31.4	42.1	21.4
0.300	33	13	22	3	64.8	91.7	37.1	40.0	18.8
0.320	33	13	22	3	64.8	91.7	37.1	40.0	18.8
0.340	33	13	22	3	64.8	91.7	37.1	40.0	18.8
0.360	33	14	21	3	66.2	91.7	40.0	38.9	17.6
0.380	31	14	21	5	63.4	86.1	40.0	40.4	26.3
0.400	31	16	19	5	66.2	86.1	45.7	38.0	23.8
0.420	31	17	18	5	67.6	86.1	48.6	36.7	22.7
0.440	30	20	15	6	70.4	83.3	57.1	33.3	23.1
0.460	29	20	15	7	69.0	80.6	57.1	34.1	25.9
0.480	27	21	14	9	67.6	75.0	60.0	34.1	30.0
0.500	26	21	14	10	66.2	72.2	60.0	35.0	32.3
0.520	22	21	14	14	60.6	61.1	60.0	38.9	40.0
0.540	21	24	11	15	63.4	58.3	68.6	34.4	38.5
0.560	19	24	11	17	60.6	52.8	68.6	36.7	41.5
0.580	19	24	11	17	60.6	52.8	68.6	36.7	41.5
0.600	17	25	10	19	59.2	47.2	71.4	37.0	43.2
0.620	17	26	9	19	60.6	47.2	74.3	34.6	42.2
0.640	17	26	9	19	60.6	47.2	74.3	34.6	42.2
0.660	15	26	9	21	57.7	41.7	74.3	37.5	44.7
0.680	13	28	7	23	57.7	36.1	80.0	35.0	45.1
0.700	8	30	5	28	53.5	22.2	85.7	38.5	48.3
0.720	8	31	4	28	54.9	22.2	88.6	33.3	47.5
0.740	7	33	2	29	56.3	19.4	94.3	22.2	46.8
0.760	4	33	2	32	52.1	11.1	94.3	33.3	49.2
0.780	3	35	0	33	53.5	8.3	100.0	0.0	48.5
0.800	2	35	0	34	52.1	5.6	100.0	0.0	49.3
0.820	1	35	0	35	50.7	2.8	100.0	0.0	50.0
0.840	1	35	0	35	50.7	2.8	100.0	0.0	50.0
0.860	1	35	0	35	50.7	2.8	100.0	0.0	50.0
0.880	0	35	0	36	49.3	0.0	100.0	.	50.7

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant des services de statistique et d'informatique de la Faculté universitaire des Sciences agronomiques et du Centre de Recherches agronomiques de Gembloux (Belgique).

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

Faculté universitaire des Sciences agronomiques
Unité de Statistique et Informatique
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLOUX (Belgique)
E-mail : statinfo@fsagx.ac.be

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

http://www.fsagx.ac.be/si/

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

PALM R. [2000]. L'analyse de la variance multivariée et l'analyse canonique discriminante : principes et applications. *Notes Stat. Inform.* (Gembloux) 2000/1, 40 p

PALM R., IEMMA A.F. [2002]. Conditions d'application et transformations de variables en régression linéaire. *Notes Stat. Inform.* (Gembloux) 2002/1, 34 p.

BROSTAUX Y. [2002]. Introduction à l'environnement de programmation statistique R. *Notes Stat. Inform.* (Gembloux) 2002/2, 22 p.

PALM R. [2003]. Le positionnement multidimensionnel : principes et application. *Notes Stat. Inform.* (Gembloux) 2003/1, 31 p.

CLAUSTRIAUX J.J. [2006]. Un regard sur les activités de l'Unité de Statistique et Informatique de la Faculté universitaire des Sciences agronomiques de Gembloux (Belgique). *Notes Stat. Inform.* (Gembloux) 2006/1, 9 p.

CARLETTI I., CLISSEN V., CLAUSTRIAUX J.J. [2006]. Introduction au logiciel Minitab sous WINDOWS. *Notes Stat. Inform.* (Gembloux) 2006/2, 23 p.

CARLETTI I., PREVOT H. [2006]. Traitement des données par le logiciel SAS : introduction au module de base. *Notes Stat. Inform.* (Gembloux) 2006/4, 31 p.

D/2006/2371/4