REVUE DE STATISTIQUE APPLIQUÉE

F. DUYME
J.-J. CLAUSTRIAUX
J.-J. DAUDIN

Qualité de validation des modèles de régression logistique binaire

Revue de statistique appliquée, tome 53, n° 3 (2005), p. 91-102
<http://www.numdam.org/item?id=RSA_2005__53_3_91_0>

Rev. Statistique Appliquée, 2005, **LIII** (3), 91-102

QUALITÉ DE VALIDATION DES MODÈLES DE RÉGRESSION LOGISTIQUE BINAIRE

F. DUYME[1], J.-J. CLAUSTRIAUX[2], J.-J. DAUDIN[3]

[1] *Institut Supérieur d'Agriculture, Lille, France*
[2] *Faculté Universitaire des Sciences Agronomiques, Gembloux, Belgique*
[3] *Institut National Agronomique, Paris-Grignon, France*

RÉSUMÉ

Dans cet article, on étudie la qualité de validation des modèles de régression logistique binaire. En particulier, on s'intéresse à l'effet de la proportion d'individus ayant le caractère étudié (codés 1) sur cette qualité. De même, on évalue l'intérêt qu'il peut y avoir à séparer ou non les étapes de sélection et d'ajustement sur des échantillons indépendants. L'ensemble de l'étude est basée sur des données simulées.

La proportion d'individus codés 1 a une influence sur la qualité des modèles. Cet effet est d'autant plus important que la proportion est très faible. Des situations de non-convergence ont d'ailleurs été rencontrées. Par contre, la qualité des modèles est assez constante pour les proportions 25 % et 50 %. Pour cette seconde modalité, la précision des estimations est un peu supérieure et surtout l'effet des autres facteurs étudiés s'estompe.

Par ailleurs, les simulations ont montré que la séparation des étapes de sélection et d'ajustement sur deux échantillons distincts ne conduit jamais à une meilleure qualité de validation des modèles. Au contraire, cette qualité est généralement moindre mais l'écart n'est que de l'ordre de 10 %.

Mots-clés : *Régression logistique binaire, proportion d'individus, nombre d'échantillons, qualité de validation.*

ABSTRACT

In this paper, we summarize a study of the quality of validation given by binary logistic regression models. We have studied the effect of the proportion of events ($i.e.$ proportion of individuals having the code 1) and the way of using data for selecting variables and fitting a model. In particular, we show the interest of performing those two steps separately or not. Artificial data were used to carry out this study.

We show that the proportion of events influences the quality of the models. Especially, this effect is important when the proportion is very low (5 % in our study). We also meet problems like non-convergence of the algorithm when trying to fit a model. However, quality is much better when a proportion equal to 25 % or 50 % is used. For the latter one, precision of the estimations is even a little bit better and the others studied factors have no effect.

The simulations also show that separating selection and fitting is not an interesting strategy because we never obtain a good quality. Difference of quality is however close to 10 per cent.

Keywords : *Binary logistic regression, proportion of individuals, number of samples, quality of validation.*

1. Introduction

Parmi les méthodes de régression, la régression logistique est employée notamment lorsque la variable dépendante ne prend que deux valeurs possibles; dans ce cas, on parle de régression logistique binaire. Celle-ci est sans aucun doute la forme la plus courante de régression logistique, dans des domaines aussi variés que la médecine, la biologie, l'économie, etc. [Collett, 1991; Ryan, 2000].

Dans ce cadre, le premier objectif de ce travail est de préciser davantage l'impact de la proportion d'individus codés 1 (le code 1 signifie la survenue de l'événement) sur la qualité de validation des modèles, sujet qui a été peu étudié, si ce n'est récemment pour le nombre d'événements par variable explicative (*EPV* ou Events Per Variable) qui est le rapport entre le nombre d'individus codés 1 et le nombre de variables explicatives. En particulier, il a été montré qu'un nombre minimum d'*EPV* doit être utilisé pour parvenir à une estimation fiable des coefficients de régression [Peduzzi *et al*., 1996; Steyergerg *et al*., 1999, 2000].

Cette question est d'importance car si on sait que la modélisation se réalise normalement lorsque la proportion d'individus dans chaque groupe est équivalente, on ne dispose guère d'information en cas de déséquilibre et a priori les résultats devraient être comparables par exemple pour 25 % ou 75 % d'individus codés 1.

Notons cependant qu'une situation de déséquilibre important est assez fréquente en médecine par exemple pour l'étude d'une maladie rare ou grave qui ne peut être basée que sur un nombre restreint d'individus dits positifs, ayant été codés 1.

Le second objectif de la recherche est relatif à la procédure de modélisation.

En effet, pour établir un modèle, trois étapes sont nécessaires : la sélection des variables explicatives, l'ajustement du modèle aux données à partir des variables retenues et la validation du modèle. Ces étapes peuvent se réaliser sur les mêmes données, ce qui est le cas si les individus sont peu nombreux; dans ce cas, on obtient un modèle « artificiellement » adéquat.

Idéalement, les étapes d'ajustement et de validation devraient être séparées c'est-à-dire réalisées sur deux échantillons distincts. Le modèle étant alors évalué sur de nouveaux individus, le biais optimiste disparaît. Ce n'est que récemment que certains auteurs ont aussi évoqué l'utilisation de trois échantillons indépendants pour réaliser les trois étapes, à savoir la sélection, l'ajustement et la validation [Celeux, 1994; Van Houwelingen et Le Cessie, 1990]. C'est pourquoi nous avons également étudié cette possibilité.

Après cette introduction (paragraphe 1), les éléments essentiels du modèle logistique sont rappelés (paragraphe 2) et nous détaillons les critères calculés. Au paragraphe 3, nous abordons la manière dont l'étude s'est déroulée, c'est-à-dire les simulations réalisées. Le paragraphe 4 est consacré à la présentation des résultats obtenus sur ces données. Une illustration sur des données réelles est fournie au paragraphe 5. Enfin, le paragraphe 6 présente les conclusions.

2. Modèle logistique

2.1. Principes

Le modèle de régression logistique fait partie d'une famille de modèles appelés modèles linéaires généralisés décrits par exemple dans les ouvrages de McCullagh et Nelder [1989] ou Dobson [1990]. Dans un modèle linéaire généralisé, la relation entre la variable à prédire Y et les variables prédictrices $X_1, \ldots, X_p$ (matrice X de p variables) est modélisée par :

$$g[E(Y)] = \boldsymbol{X\beta},$$

où $E(Y)$ est l'espérance mathématique de la variable aléatoire Y, $\boldsymbol{X\beta} = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p$ avec $\beta_0, \ldots, \beta_p$ les coefficients, et g est une fonction appelée fonction de lien. La fonction g adaptée lorsque Y est une variable binaire peut-être la fonction $logit, probit, log(-log)$. D'après Collett [1991], la fonction $logit$ est la plus employée pour sa simplicité essentiellement.

On définit, pour un individu i, la probabilité π_i (on parle aussi de probabilité *a posteriori* ou probabilité de l'événement) d'être dans le groupe des individus codés 1. La forme du modèle logistique s'écrit :

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

où β_0 et β_1 sont les coefficients du modèle et x_i la i$^{\text{ème}}$ valeur de la variable explicative X. 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 le log représente le logarithme népérien. Lorsque p variables sont disponibles, la fonction g s'écrit (avec j l'indice des variables et i celui des individus) :

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

ou

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

2.2. Validation externe

Lorsqu'un nombre suffisant d'individus sont disponibles, alors il est courant d'en réserver une partie pour l'ajustement du modèle, l'autre partie servant pour la validation. On parle ainsi respectivement d'échantillon d'apprentissage et d'échantillon

de validation. La proportion d'individus réservés pour la validation est couramment de 25 % à 50 % du nombre total d'individus.

Les coefficients du modèle, obtenus sur le premier échantillon, sont alors utilisés sur le second échantillon de manière à calculer des probabilités *a posteriori* $\widehat{\pi}_i$. La qualité de validation est ensuite déterminée sur la base de ces probabilités et des valeurs binaires y_i où y_i prend la valeur 1 si l'événement d'intérêt a lieu, 0 sinon.

Pour ce travail, deux critères sont considérés : la statistique de Hosmer et Lemeshow et le critère c, encore connu sous l'appellation surface sous la courbe *ROC (Receiver Operating Characteristic)*.

La statistique de Hosmer et Lemeshow [1989], notée χ^2_{HL}, équivaut à une analyse des résidus; elle est basée sur le regroupement des individus en k classes de taille presque identique (10 classes dans cette étude) :

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

où n_k est la taille de la classe k, o_k est le nombre d'individus codés 1 de la classe et e_k est dans ce cas la fréquence attendue de cette classe, c'est-à-dire la somme des probabilités *a posteriori*. Le nombre de degrés de liberté de la variable χ^2 est dans ce cas égal au nombre de classes.

Le second critère mesure les capacités du modèle à classer de nouveaux individus dans la bonne catégorie (code 0 ou 1). Il s'agit du critère c [SAS, 1995] :

$$c = \frac{n_c + 0,5(t - n_c - n_d)}{t},$$

où n_c est le nombre de paires concordantes, n_d le nombre de paires discordantes, t le nombre total de paires c'est-à-dire $t = n(n-1)/2$ et n l'effectif total. La quantité entre parenthèses représente le nombre de paires d'*ex æquo*. Si elle est négligeable, alors le critère c est approximativement égal à la proportion de paires concordantes.

2.3. La non-convergence

La méthode d'ajustement d'un modèle de régression logistique aux données disponibles est celle du maximum de vraisemblance. Pour arriver à maximiser la fonction de vraisemblance, il est nécessaire d'utiliser un algorithme particulier (procédure itérative) décrit par exemple dans McCullagh et Nelder [1989] ou Collett [1991].

Il n'est cependant pas toujours possible de déterminer avec précision une estimation des coefficients du modèle. Ainsi, dans le cas de la séparation complète des deux groupes d'individus codés 0 ou 1 dans l'espace des variables explicatives, l'algorithme ne parvient pas à converger, c'est-à-dire qu'il ne permet pas d'obtenir une estimation précise des coefficients du modèle (estimation et erreur standard avec des valeurs très grandes le plus souvent). Dans ce cas, on parle de non-convergence et certains logiciels comme SAS ou MINITAB indiquent ce problème.

3. Simulations

3.1. Facteurs étudiés

L'étude concerne, d'une part, la proportion d'individus codés 1 (n_1/n), à savoir 5 %, 25 % et 50 % et, d'autre part, la manière d'utiliser les données pour mettre en place les modèles et les valider. Pour la validation, un échantillon est isolé. Pour établir les modèles sur les données restantes, nous définissons deux enchaînements (*ench*) c'est-à-dire deux manières d'utiliser les données :

- type 1-1 : toutes les données restantes servent à sélectionner et à ajuster un modèle;
- type 1-2 : la sélection a lieu sur la moitié des données restantes, et l'ajustement sur l'autre moitié.

Par ailleurs, deux autres facteurs sont envisagés : le nombre de variables explicatives et l'écart-type de la part aléatoire ou bruit.

Le nombre de variables explicatives retenues $(nVar)$ est 6 ou 12, avec pour chaque cas uniquement trois variables en relation directe avec le vecteur Y (variables utiles); ce sont les variables X_1, X_2, X_3 qui servent à définir Y (paragraphe 3.2).

Pour l'écart-type (*seps*) de la part aléatoire que l'on ajoute au modèle lors de la génération des données (paragraphe 3.2.), trois valeurs sont retenues : 0,5, 2 et 3,5. Ces valeurs ont été choisies, lors d'essais préliminaires, de manière à avoir respectivement une très bonne, une bonne et une moyenne qualité d'ajustement.

3.2. Génération des données

12 variables indépendantes de distribution uniforme $U(0;1)$ sont générées. Les trois premières variables définissent le vecteur Y de la manière suivante :

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3,$$

où b_j $(j = 0, \ldots, 3)$ sont les coefficients dont les valeurs valent respectivement 0, 4, 2, 1. Ces valeurs sont arbitraires, mais résultent de travaux préliminaires sur données observées qui ont montré que X_1 est très discriminante, X_2 un peu moins et X_3 encore moins. Soit ε le vecteur représentant la part aléatoire. Ce vecteur est généré selon une loi $N(0; seps)$ avec *seps* l'écart-type valant 0,5, 2 ou 3,5. Les vecteurs Y et ε sont alors ajoutés pour définir un nouveau vecteur Yc :

$$Yc = Y + \varepsilon.$$

On procède alors au codage des valeurs de Yc en 0 ou 1, selon la proportion n_1/n utilisée (5 %, 25 % ou 50 %). Les $n - n_1$ plus petites valeurs du vecteur Yc reçoivent le code 0 et les n_1 autres valeurs le code 1.

Sur l'ensemble des individus générés, 10 000 sont réservés pour la validation. Les autres sont répartis en échantillons de taille 400 (enchaînement de type 1-1) ou $200 + 200$ (type 1-2) de manière aléatoire et sans remise. La taille de l'échantillon de validation est assez importante de manière à obtenir une estimation fiable d'une proportion d'individus mal classés (surtout dans le cas où $n_1/n = 5$ %).

Pour chaque combinaison des facteurs étudiés, 20 répétitions ont été prévues.

La méthodologie adoptée est inspirée des travaux de Derksen et Keselman [1992], Hosmer *et al*. [1997] et Press et Wilson [1978]. Les premiers ont étudié l'impact d'un mélange de variables utiles et inutiles sur les résultats d'un processus de sélection automatique en régression linéaire multiple. Les seconds justifient l'emploi de valeurs préalablement fixées pour les coefficients d'un modèle de régression logistique, tandis que les derniers, dans leur méthodologie de traitement des données, ont rendu binaire une variable au préalable continue.

Notons enfin que toutes les sélections automatiques de variables et les ajustements ont été effectués à l'aide du logiciel SAS.

4. Analyse des résultats

4.1. Problème de la non-convergence

La méthode *stepwise* de sélection des variables qui a été employée nécessite des ajustements de modèle aux données de l'échantillon de sélection. Il est donc tout à fait possible que des cas de non-convergence surviennent.

Une non-convergence pour une ou plusieurs variables retenues dans le cas d'un enchaînement de type 1-1 signifie un abandon du modèle car les estimations des coefficients fournies sont inexploitables. Pour un enchaînement de type 1-2, nous distinguons la sélection de l'ajustement. Nous avons ainsi relevé les cas de non-convergence en sélection. Les variables proposées ont tout de même été utilisées en ajustement de manière à «récupérer» le maximum de modèles.

Comme le montre le tableau 1, une non-convergence est principalement constatée lorsque très peu d'individus sont codés 1 (proportion n_1/n de 5 %) et que l'écart-type *seps* est petit (0,5). Signalons que pour des échantillons de taille 400, 20 individus sont codés 1, ce qui représente un *EPV* de 3,3 ou 1,7 selon le nombre de variables explicatives (6 ou 12). D'après les travaux de Peduzzi *et al* [1996] et Steyerberg *et al*. [1999, 2000], un *EPV* minimum de 10 est à respecter.

TABLEAU 1

Proportions de cas de non-convergence en ajustement en fonction des facteurs étudiés (déterminées par rapport au nombre de répétitions); en grisé : aucune non-convergence rencontrée

	seps	*0,5*			*2*			*3,5*		
	n_1/n	*5%*	*25%*	*50%*	*5%*	*25%*	*50%*	*5%*	*25%*	*50%*
ench	*nVar*									
1-1	6	17%								
	12	52%								
1-2	6	27%	2%							
	12	25%	7%		2%					

En dehors des données de ce tableau, nous avons aussi pu remarquer des cas de presque non-convergence, détectables par des fortes valeurs des estimations $\widehat{b}_j$ des coefficients b_j et/ou de l'écart-type de ces estimations. Une détection visuelle de ces grandes valeurs des écarts-types des coefficients a été confirmée par le calcul du déterminant des matrices de variances-covariances (matrices associées aux estimateurs des paramètres) dont les valeurs s'échelonnent de 10^{-4} à environ $3 \cdot 10^5$. Ainsi, pour la détermination des critères c et χ^2_{HL}, des modèles ont été éliminés, le calcul des probabilités *a posteriori* n'ayant plus aucun sens (déterminant supérieur à 250).

4.2. Analyse des valeurs de χ^2_{HL}

Parmi les facteurs étudiés, le type d'enchaînement représente la source de variation la plus importante. L'enchaînement de type 1-2 donne en moyenne des valeurs 12 % plus élevées que celles du type 1-1. Le nombre de variables est le second facteur de variation des résultats. Lorsque 12 variables sont employées, les valeurs sont nettement plus élevées. L'écart-type de la part aléatoire conduit également à des résultats très variables. Il s'agit surtout de la première modalité ($seps = 0,5$) qui procure des valeurs de χ^2_{HL} les plus faibles, alors que les deux autres modalités donnent des valeurs assez similaires. Enfin, parmi les modalités du facteur n_1/n, c'est la proportion 25 % qui permet d'obtenir les plus faibles valeurs de χ^2_{HL}.

Ces commentaires sont illustrés par le graphique de la figure 1 sauf en ce qui concerne le facteur $nVar$.

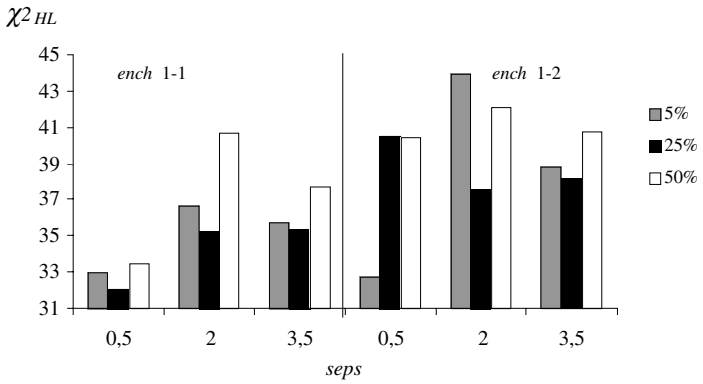

FIGURE 1
Valeurs moyennes du χ^2_{HL} en fonction du type d'enchaînement (ench), de l'écart-type de la part aléatoire (seps) et de la proportion d'individus codés 1 (5 %, 25 %, 50 %)

Tenant compte des problèmes de non-convergence évoqués au paragraphe 4.1, si on retire de l'analyse la modalité 5 % du facteur n_1/n et la modalité 0,5 du facteur *seps*, les valeurs de χ^2_{HL} sont dans ces conditions beaucoup moins sujettes à des variations. Il en ressort néanmoins que (tableau 2) :

– l'enchaînement de type 1-2 donne systématiquement des valeurs plus élevées et donc une moins bonne qualité de validation. Cependant, la progression de χ^2_{HL} entre les deux types d'enchaînement est toujours inférieure à 10 %;
– la proportion 25 % de n_1/n conduit à des résultats un peu meilleurs à ceux de 50 %; dans ce cas ($n_1/n = 50$ %), l'effet du nombre de variables explicatives est plus important;
– le nombre de variables explicatives n'est pas un facteur de variation lorsque $n_1/n = 50$ %.

TABLEAU 2
Valeurs de χ^2_{HL} (entre parenthèses : écart-type) pour deux des trois modalités des facteurs seps et n_1/n

	n_1/n	25 %		50 %	
	$nVar$	6	12	6	12
seps	*ench*				
2	1-1	34,4 (6,6)	36,1 (6,2)	40,6 (3,2)	40,9 (2,7)
	1-2	37,2 (7,9)	38,0 (8,4)	42,2 (4,2)	42,3 (4,6)
3,5	1-1	34,5 (5,3)	36,2 (5,2)	38,2 (3,1)	37,3 (3,3)
	1-2	37,4 (6,2)	39,1 (5,2)	40,9 (5,5)	40,5 (5,7)

4.3. Analyse des valeurs de c

Les deux facteurs de variation les plus importants sont l'écart-type de la part aléatoire et la proportion d'individus codés 1. Sans surprise, quand l'écart-type *seps* augmente, les valeurs de c diminuent. Pour le second facteur, la progression est également inversement proportionnelle. La modalité 5 % conduit donc aux plus fortes valeurs de c donc à la meilleure qualité prédictive.

En ce qui concerne le nombre de variables explicatives, la modalité 12 aurait tendance à donner des valeurs plus faibles pour c donc une moins bonne qualité de prédiction. Ce constat est également vérifié pour l'enchaînement 1-2 par rapport au type 1-1 (figure 2).

Si, comme pour le critère χ^2_{HL}, nous excluons de l'analyse la première modalité des facteurs *seps* et n_1/n, l'écart des valeurs de c entre les deux types d'enchaînement est toujours inférieur à 1 % quels que soient $nVar$, *seps* et n_1/n, ce qui évidemment est insignifiant. Notons enfin que l'effet du nombre de variables explicatives est plus faible lorsque 50 % d'individus codés 1 sont utilisés.

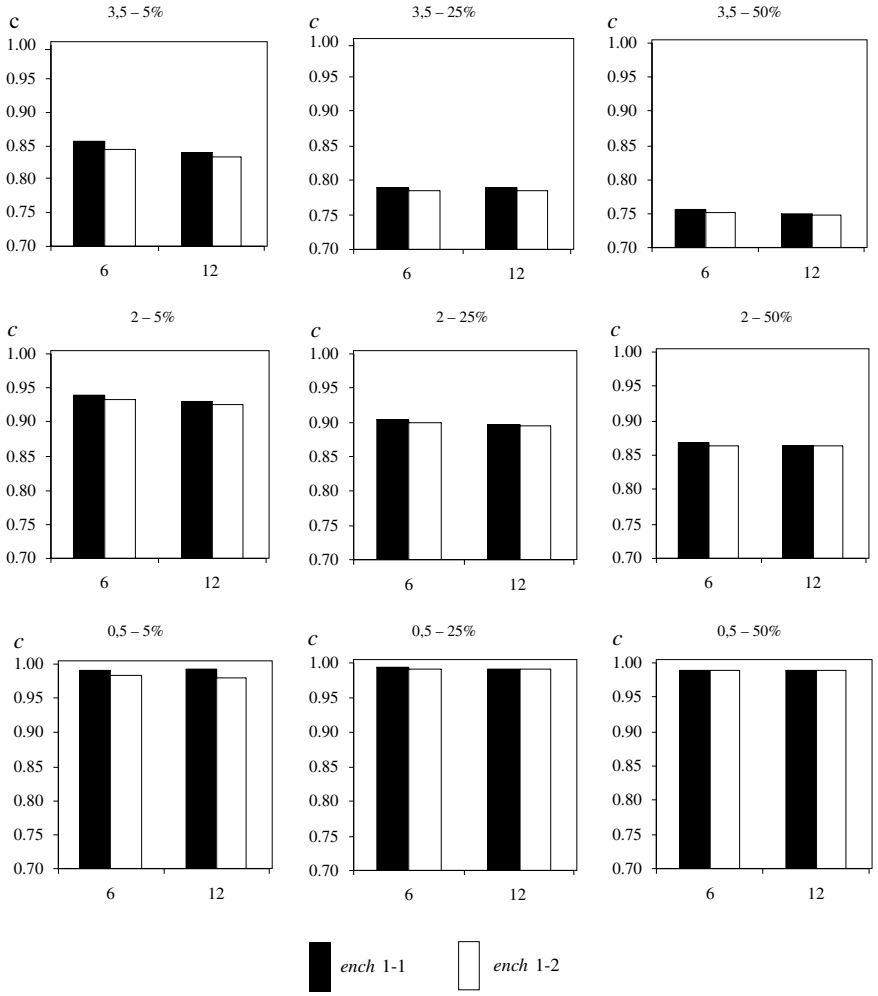

FIGURE 2
Valeurs moyennes de c pour chaque combinaison $seps - n_1/n$ en fonction du type d'enchaînement (ench) et du nombre de variables explicatives (6 ou 12)

5. Application à des données réelles

Le travail sur données simulées a été poursuivi sur quelques jeux de données réelles [Duyme, 2001]. Pour cela, trois jeux de données ont été recueillis : il s'agit de données agronomiques relatant le parcours ou itinéraire cultural de trois cultures communes dans le nord de la France (la chicorée, la betterave sucrière et

le blé). La variable à expliquer est le rendement de chaque culture exprimé en unité conventionnelle par hectare. L'individu statistique est dans cette étude la parcelle. Les variables explicatives présentent peu voire très peu de corrélation : aucune valeur de *VIF* (Variance Inflation Factor) ne dépasse 3 quelle que soit la culture.

Pour chaque culture, nous disposions d'environ 375 parcelles, scindées en trois échantillons de même taille (un échantillon par étape de construction d'un modèle – sélection, ajustement et validation). La variable Y (le rendement) a été rendue binaire en utilisant deux des trois quartiles, permettant de repérer 25 % ou 50 % des parcelles ayant le meilleur rendement. La qualité de validation a été déterminée par la statistique de Hosmer et Lemeshow, χ^2_{HL}, en forçant une répartition en 8 classes, afin d'avoir suffisamment d'individus par classe.

Les résultats moyens sont donnés à la figure 3. Ainsi, pour chaque culture, nous remarquons que l'enchaînement de type 1-2 fournit des valeurs de la statistique de Hosmer et Lemeshow un peu ou très supérieures à celles de l'autre enchaînement.

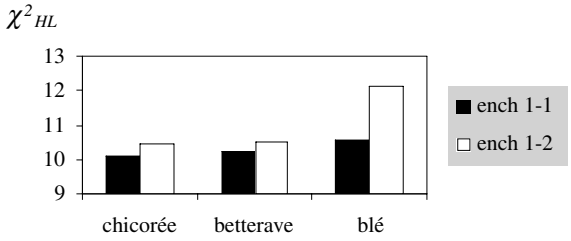

FIGURE 3
Valeurs moyennes du χ^2_{HL} en fonction du type d'enchaînement (ench), par culture

6. Conclusion

Le principal problème survenu dans cette étude par simulation est la non-convergence, essentiellement pour une faible valeur d'écart-type (*seps*) et pour une proportion de valeurs codées 1 voisine de 5 %. La non-convergence ne permet pas d'ajuster un modèle car les estimations des coefficients et des écarts-types sont en dehors de toute réalité. La plupart des logiciels affichent à ce titre un message d'avertissement.

Par ailleurs, il existe des situations où l'algorithme d'ajustement parvient à converger, mais avec néanmoins une mauvaise précision des estimations des coefficients. C'est la raison pour laquelle nous avons préconisé le calcul du déterminant de la matrice de variances-covariances associée aux estimateurs des paramètres. Ce calcul devrait être systématiquement réalisé lorsque seulement 5 % seulement des individus sont codés 1 ($n_1/n = 5$ %). Un moyen d'éviter le problème de non-convergence ou de presque non-convergence consiste à ré-équilibrer le nombre d'individus codés 0 par rapport aux individus de l'autre code : cela reviendrait à échantillonner de manière aléatoire et sans remise parmi les individus au code 0 afin de diminuer leur nombre et donc de diminuer la taille de l'échantillon, ce qui permet d'accroître la valeur de

n_1/n. Ainsi, par exemple, sur 500 individus dont seulement 25 seraient codés 1, cela reviendrait, si on désire avoir 25 % de 1, à n'utiliser que 20 % de l'effectif total, ce qui est difficilement recommandable. Si par contre la taille totale de l'échantillon ne permet pas de réduire le nombre d'individus codés 0 de manière importante, alors nous conseillons soit d'avoir quand même plus de 5 % d'individus codés 1 soit d'envisager l'utilisation de la régression logistique exacte qui dans ce cas permet de traiter tous les individus de l'échantillon.

En dehors de ces situations spéciales, nous avons remarqué que la modalité 25 % de n_1/n aboutit à une qualité de validation des modèles supérieure à celle obtenue pour 50 %. Néanmoins, la précision des estimations est moins bonne. Ainsi, pour un enchaînement du type 1-2, nous recommandons d'équilibrer les nombres de 0 et de 1. Passer de 25 % à 50 % d'individus codés 1 revient à ne prendre que la moitié des données disponibles. Pour un enchaînement du type 1-1, l'effet du facteur $nVar$ est insignifiant.

Toujours pour un écart-type de l'erreur (*seps*) valant 2 ou 3,5, mais pour $n_1/n = 50\,\%$, la précision des estimations des coefficients des modèles est très bonne, surtout pour un enchaînement du type 1-1. La qualité de validation est dans ce cas très peu dépendante du type d'enchaînement et du nombre de variables explicatives. Nous préconisons cependant d'utiliser un enchaînement du type 1-1.

En résumé, un enchaînement du type 1-2 n'est jamais souhaitable. Cela signifie qu'il n'est pas conseillé de séparer la sélection de l'ajustement sur des échantillons deux fois plus petits. Il complique les manipulations de données et n'apporte pas une meilleure précision des estimations ni une meilleure qualité de validation. Cet enchaînement devient équivalent (mais pas meilleur) à l'autre type (1-1) à condition que *seps* soit grand (3,5 voire plus, sans que nous ne puissions donner d'information précise pour $seps > 3,5$).

Notons pour terminer que les manipulations sur données réelles aboutissent à des observations similaires à ce que nous avons obtenu sur données simulées.

Remerciements. – L'accomplissement de ce travail a été facilité par les nombreux conseils et les remarques constructives du Professeur J.K. Lindsey (Université de Liège) et du Professeur R. Palm (Faculté Universitaire des Sciences Agronomiques de Gembloux). Nous tenons à les en remercier.

Références

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

COLLETT D. (1991), *Modelling binary data*. London, Chapman & Hall, 369p.

DERKSEN S., KESELMAN H.J. (1992), Backward, forward and stepwise automated subset selection algorithms : frequency of obtaining authentic and noise variables. *British J. Math. Stat. Psych*. **45**, 265-282.

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

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.

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

HOSMER D.W., HOSMER T., LE CESSIE S., LEMESHOW S. (1997), A comparison of goodness-of-fit tests for the logistic regression model. *Stat. Med.* **16**, 965-980.

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

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. (2000), Some issues in logistic regression. Comm. *Stat.-Theo. Meth.* **29**, 2019-2032.

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

STEYERBERG E.W., EIJKEMANS M.J.C., HABBEMA J.D.F. (1999), Stepwise selection in small data sets : a simulation study of bias in logistic regression analysis. *J. Clin. Epidemiol.* **52**, 935-942.

STEYERBERG E.W., EIJKEMANS M.J.C., HARRELL F.E., HABBEMA J.D.F. (2000), Prognostic modelling with logistic regression analysis : a comparison of selection and estimation methods in small data sets. *Stat. Med.* **19**, 1059-1079.

VAN HOUWELINGEN J.C., LE CESSIE S. (1990), Predictive value of statistical models. *Stat. Med.* **9**, 1303-1325.