B A
S E

Biotechnol. Agron. Soc. Environ. 2011 **15**(3), 425-433

Le **Point** *sur :*

Principaux modèles utilisés en régression logistique

Adeline Gillet [(1)], Yves Brostaux [(2)], Rodolphe Palm [(2)]

[(1)] Centre de Recherche Public Gabriel Lippmann. Département Environnement et Agro-biotechnologies. Rue du Brill, 41. L-4422 Belvaux (Grand-Duché du Luxembourg).
[(2)] Université de Liège. Gembloux Agro-Bio Tech. Unité de Statistique, Informatique et Mathématique appliquées. Avenue de la Faculté d'Agronomie, 8. B-5030 Gembloux (Belgique). E-mail : Rodolphe.Palm@ulg.ac.be

Reçu le 18 janvier 2010, accepté le 16 novembre 2010.

La régression est une technique très couramment utilisée pour décrire la relation existant entre une variable à expliquer et une ou plusieurs variables explicatives. Lorsque la variable à expliquer est une variable qualitative, la régression linéaire classique au sens des moindres carrés doit être abandonnée au profit de la régression logistique. Si la variable à expliquer ne présente que deux modalités, on utilise la régression logistique binaire. Si elle présente plus de deux modalités et si celles-ci ne sont pas ordonnées, on doit employer la régression logistique polychotomique nominale. Enfin, si la variable à expliquer présente plus de deux modalités et que celles-ci sont ordonnées, la méthode à exploiter est la régression polychotomique ordinale. Cette note décrit ces trois méthodes de régression logistique et, pour la régression ordinale, elle présente les trois modèles les plus souvent utilisés. Ces modèles sont illustrés par un exemple relatif au dépérissement du chêne en Région wallonne (Belgique).
Mots-clés. Logistique, régression, binaire, nominal, ordinal, polychotomique, méthode statistique, modèle.

Main models used in logistic regression. Regression is a commonly used technique for decribing the relationship between a response variable and one or more explanatory variables. When the response variable is a categorical variable, usual regression based on ordinary least squares should be replaced by logistic regression. Binary logistic regression should be used to perform a regression on a dichotomous response. Nominal polytomous logistic regression applies to a categorical response variable that has more than two levels with no natural ordering. And ordinal polytomous logistic regression is used when the response is a categorical variable that has more than two levels with a natural ordering. This note gives an overview of these logistic regression methods and describes three models commonly used when performing ordinal logistic regression. These models are illustrated by an example related to oak decline in the Walloon Region (Belgium).
Keywords. Logistic, regression, binary, nominal, ordinal, polytomous, statistical method, model.

1. INTRODUCTION

La régression linéaire, simple ou multiple, est une méthode statistique très couramment utilisée dans le 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 x_1, x_2, ..., x_p, appelées prédicteurs. La méthode est cependant limitée aux situations où la variable à expliquer est une variable quantitative dont la distribution, pour une valeur fixée des prédicteurs, est normale. Elle ne devrait notamment pas être utilisée lorsque la variable y est une variable qualitative. Pour de telles situations, la méthode indiquée est la régression logistique qui offre plusieurs variantes en fonction du nombre et de la nature des classes de la variable à expliquer.

La première méthode, appelée régression logistique binaire (*binary logistic regression*), correspond au cas où la variable y comporte uniquement deux classes, les individus étant décrits par la présence ou l'absence d'un caractère donné. Par exemple, des individus (parcelles, plantes, animaux, etc.) peuvent être attaqués ou non par un parasite, être fertiles ou non, être porteurs ou non d'une tare, etc.

La deuxième méthode, appelée régression logistique polychotomique nominale (*polytomous nominal logistic regression*), permet de traiter les cas où la variable à expliquer possède plus de deux classes si celles-ci ne peuvent pas être ordonnées ou si on ne souhaite pas tenir compte de l'ordre dans le cas où elles seraient ordonnées. Une telle situation se présente par exemple si des individus sont caractérisés par l'appartenance à une espèce donnée, par une couleur ou par le choix d'une réponse à une question posée

parmi trois propositions telles que « oui », « non », « ne sait pas ».

Enfin, la troisième méthode, appelée régression polychotomique ordinale (*polytomous ordinal regression*), concerne les situations où la variable *y* présente plus de deux modalités qui peuvent être ordonnées et dont on souhaite tenir compte de l'ordre. Un exemple typique est la description de l'intensité de l'attaque d'individus par un parasite, cette description étant réalisée par exemple sur la base d'une échelle à quatre niveaux notés A, B, C et D, le niveau A représentant l'absence d'attaque, le niveau B une attaque faible, le niveau C une attaque modérée et le niveau D une attaque forte.

Dans la pratique, il arrive fréquemment que l'appartenance aux classes soit décrite par des codes numériques. Ainsi, au lieu de noter les degrés d'attaque par A, B, C ou D, on peut les identifier par les codes 0, 1, 2 ou 3. Ce codage est purement arbitraire et les méthodes logistiques ne considèrent jamais les valeurs numériques en tant que telles, mais simplement comme des noms de modalités. Dans le cas d'une variable ordinale codée, il est d'ailleurs possible que l'ordre logique des classes ne corresponde pas à l'ordre donné par les nombres utilisés comme codes. Une telle situation est cependant à déconseiller car elle est de nature à perturber inutilement l'interprétation des résultats.

Pour les trois méthodes citées ci-dessus, le but est de modéliser une ou plusieurs probabilités liées à l'appartenance aux classes, en fonction d'un ou de plusieurs prédicteurs, qui peuvent eux-mêmes être des variables quantitatives ou des variables qualitatives, supposées parfaitement connues.

L'objectif de cette publication est de présenter ces méthodes, de manière succincte, en insistant sur l'interprétation des modèles.

Après cette introduction, nous examinons d'abord le modèle de régression binaire, puis le modèle polychotomique nominal. Nous passons ensuite en revue trois modèles pour la régression polychotomique ordinale. Nous clôturons par une discussion relative aux modèles polychotomiques.

Dans un but de simplification, les différents modèles sont présentés dans le cas de la régression sur un seul prédicteur, noté *x*. La généralisation au cas de plusieurs prédicteurs est immédiate : il suffit, dans les notations des modèles, de remplacer le coefficient de régression et la variable explicative par des vecteurs de coefficients et de variables.

Nous illustrons les différents modèles par un exemple traité antérieurement par Gillet (2005 ; 2007) et relatif au dépérissement de chênes dans le Condroz et l'Ardenne belge. Les données utilisées concernent le niveau de dépérissement de 230 chênes et l'altitude des stations dans lesquelles ces chênes ont été observés. Le dépérissement a été évalué par l'aspect du houppier sur une échelle à quatre niveaux. Le niveau 1 correspond à un dépérissement très faible, le niveau 2 à un dépérissement faible, le niveau 3 à un dépérissement fort et le niveau 4 à un dépérissement très fort. La variable à expliquer *y* est donc une variable qualitative ordinale à quatre classes et la variable explicative *x* est l'altitude. Des modifications dans ces variables sont cependant apportées afin de permettre l'illustration des différentes situations. Les modèles qui sont ajustés aux données servent uniquement d'exemples et ne représentent pas nécessairement des modèles adéquats pour la modélisation du dépérissement, celui-ci étant lié à d'autres facteurs que la seule altitude.

On notera que la régression logistique soulève bien d'autres problèmes que la définition du modèle : problèmes d'inférence statistique, de critères de qualité de l'ajustement et de choix de variables explicatives, notamment. Ces aspects ne sont pas abordés dans cette note. Des informations complémentaires à ce sujet sont données dans les ouvrages consacrés, totalement ou partiellement, à la régression logistique. Parmi ceux-ci, nous citerons les livres d'Agresti (2002) et Hosmer et al. (2000). Pour la régression binaire, des informations sont également données dans le document de Duyme et al. (2006).

2. MODÈLE LOGIT POUR DONNÉES BINAIRES

Lorsque la variable à expliquer possède deux modalités codées par exemple $y = 1$ et $y = 2$, l'objectif est de modéliser, en fonction de x, la probabilité d'appartenance à une des deux catégories, appelée succès ou évènement (*success* ou *event*). Nous notons cette probabilité $\pi(x_i)$, ou plus simplement π.

Les probabilités $\pi(x_i)$ évoluent cependant de manière non linéaire en fonction de x_i. De plus, la variance de ces probabilités varie avec x_i. Il en résulte que l'utilisation d'un modèle linéaire exprimant π en fonction de x et ajusté par les moindres carrés classiques n'est pas une solution adéquate, les conditions d'application de la régression – linéarité et constance de la variance conditionnelle – n'étant pas remplies. Pour cette raison, on effectue une transformation de la probabilité de succès $g(\pi(x_i))$. Cette transformation s'appelle fonction de lien (*link function*), et par la suite sera notée simplement g.

Plusieurs fonctions de lien existent mais la plus couramment utilisée est la fonction logit :

$$g = \text{logit}(\pi) = \log_e [\pi / (1 - \pi)],$$

parce qu'elle conduit à une interprétation simple des résultats, mais aussi pour des raisons théoriques (Collett, 1999).

Le modèle de régression s'écrit alors :

$$g = \alpha + \beta\, x,$$

où α et β sont des paramètres à estimer, le plus souvent par la méthode du maximum de vraisemblance. La transformation inverse permet alors de retrouver les probabilités estimées en fonction de x :

$$\pi = \exp(g) / [1 + \exp(g)],$$

qui sont toujours comprises entre 0 et 1.

Pour une valeur x_i donnée, le rapport entre la probabilité de succès π et la probabilité d'échec $1 - \pi$ est appelé chance ou cote, mais est le plus souvent désigné par le terme anglais *odds*. Il est égal à :

$$\pi / (1 - \pi) = \exp(g).$$

Lorsque la probabilité de succès est plus grande que la probabilité d'échec, l'*odds* est supérieur à l'unité. Si les deux probabilités sont égales, l'*odds* est égal à 1. Enfin, si la probabilité de succès est plus petite que la probabilité d'échec, l'*odds* est inférieur à l'unité.

Si on considère maintenant le rapport entre les *odds* relatifs à $x_i + 1$ et à x_i, on définit le rapport de chances ou le rapport des cotes, plus souvent désigné par le terme anglais *odds ratio*, qui est directement lié au coefficient de régression β :

$$\frac{\pi(x_i + 1) / [1 - \pi(x_i + 1)]}{\pi(x_i) / [1 - \pi(x_i)]} = \exp(\beta).$$

L'*odds ratio* n'est jamais négatif, mais n'a pas de borne supérieure. Une valeur égale à l'unité signifie que la cote pour x_i est égale à la cote pour $x_i + 1$. Dans cette situation, la variable explicative n'a donc pas d'effet sur la cote et le coefficient de régression est nul. Un *odds ratio* inférieur à l'unité correspond à un coefficient de régression négatif et signifie que la probabilité de succès diminue lorsque x augmente. Un *odds ratio* supérieur à l'unité correspond à un coefficient de régression positif et signifie que la probabilité de succès augmente lorsque x augmente.

Lorsque la variable explicative est continue, l'*odds ratio* est parfois très proche de 1, une différence d'une unité de x étant insuffisante pour modifier de manière sensible les rapports des cotes. Dans ce cas, il peut être préférable de calculer l'*odds ratio* pour une modification δ de la variable explicative. On a alors :

$$\frac{\pi(x_i + \delta) / [1 - \pi(x_i + \delta)]}{\pi(x_i) / [1 - \pi(x_i)]} = \exp(\delta\beta).$$

Pour illustrer la régression logistique binaire à partir des données relatives au dépérissement du chêne, nous regroupons les classes pour lesquelles le code est supérieur à 1. Après ce regroupement, la première classe correspond au dépérissement très faible et la deuxième correspond au dépérissement faible à très fort. L'appartenance à la première classe est choisie comme l'évènement. Le modèle ajusté est le suivant :

$$g = 1{,}5714 - 0{,}009533\, x.$$

La valeur négative du coefficient de régression indique que la probabilité d'être dans la classe « dépérissement très faible » diminue avec l'altitude. Si on détermine, à titre d'exemple, cette probabilité pour une altitude de 200 m et pour une altitude de 300 m, en utilisant la relation donnée ci-dessus, on trouve :

$$\pi(200) = 0{,}417 \text{ et } \pi(300) = 0{,}216.$$

Ces valeurs correspondent à des probabilités estimées par le modèle de régression. Elles devraient donc être notées :

$$\hat{\pi}(200) \text{ et } \hat{\pi}(300).$$

Toutefois, pour alléger les notations, nous choisissons de représenter systématiquement les valeurs théoriques et les valeurs estimées par le même symbole, le contexte permettant de lever toute ambigüité. À partir de ces probabilités, on vérifie bien que, pour $\delta = 100$, l'*odds ratio* est égal à :

$$\frac{\pi(300)/[1 - \pi(300)]}{\pi(200)/[1 - \pi(200)]} = \exp[(100)(-0{,}009533)] = 0{,}385.$$

Lorsque l'altitude augmente de 100 m, le rapport entre la probabilité d'un dépérissement très faible et un dépérissement faible à très fort est donc multiplié par 0,39 ou encore divisé par 2,6 environ.

Le **tableau 1** donne, en fonction de l'altitude, les probabilités π d'être dans la classe de dépérissement

Tableau 1. Données binaires : probabilité d'un dépérissement très faible (π), probabilité d'un dépérissement faible à très fort ($1 - \pi$) et *odds* [$\pi / (1 - \pi)$] en fonction de l'altitude — *Binary data: probability of a very slight decline* (π), *probability of a decline from slight to very strong* ($1 - \pi$) *and odds* [$\pi / (1 - \pi)$] *as a function of altitude*.

Altitude	π	$1 - \pi$	$\pi / (1 - \pi)$
200	0,417	0,583	0,715
300	0,216	0,784	0,275
400	0,096	0,904	0,106
500	0,039	0,961	0,041
600	0,016	0,984	0,016

très faible, les probabilités 1 - π d'être dans la classe faible à très fort, ainsi que les rapports de ces deux probabilités qui sont les *odds*. On vérifie bien ainsi que le rapport entre un *odds* et l'*odds* précédent est constant et égal, aux erreurs d'arrondis près, à 0,385.

3. MODÈLE LOGIT POUR DONNÉES NOMINALES

La régression logistique polychotomique nominale est une extension naturelle de la régression binaire qui permet de prendre en compte un nombre de catégories supérieur à deux. Soit $y = j$, l'indication de l'appartenance d'un individu à la catégorie j. Une catégorie est choisie comme référence et l'appartenance à cette catégorie est considérée comme l'évènement de référence (*reference event*). Pour une variable à expliquer à k modalités et en prenant arbitrairement la k^e modalité comme référence, on modélise l'évolution, en fonction de x, de k-1 probabilités conditionnelles :

$$\omega_j = P(y = j \,|\, y = j \text{ ou } y = k) = \pi_j / (\pi_j + \pi_k), (j = 1,..., k - 1).$$

Les logits correspondant à ces probabilités sont appelés logits pour une catégorie de référence (*baseline-category logits*) :

$$g_j = \log_e [\omega_j / (1 - \omega_j)] = \log_e (\pi_j / \pi_k)$$

et sont modélisés en fonction de x par les relations suivantes : $g_j = \alpha_j + \beta_j x$.

Un résultat similaire pourrait être obtenu en ajustant indépendamment k - 1 régressions binaires, chacune de ces régressions étant ajustée au sous-ensemble d'individus appartenant à la catégorie j et à la catégorie k. Les ajustements séparés donneraient cependant des paramètres estimés légèrement différents et moins efficaces que l'ajustement simultané des k - 1 équations.

Disposant des k - 1 équations, on peut déterminer les *odds* : $\pi_j / \pi_k = \exp(g_j)$ et retrouver les probabilités d'appartenance à chacune des classes par les relations suivantes :

$$\pi_j = \exp(g_j) / \left[1 + \sum_{l=1}^{k-1} \exp(g_l)\right]$$

et $$\pi_k = 1 - (\pi_1 + ... + \pi_{k-1}) = 1 / \left[1 + \sum_{l=1}^{k-1} \exp(g_l)\right].$$

Les coefficients de régression β_j sont liés aux *odds ratios* :

$$\frac{\pi_j(x_i + \delta) / \pi_k(x_i + \delta)}{\pi_j(x_i) / \pi_k(x_i)} = \exp(\delta\beta_j).$$

Il en résulte qu'un coefficient de régression positif signifie que l'*odds* est plus grand en $x_i + \delta$ qu'en x_i ou encore que la probabilité d'appartenance à la catégorie j augmente plus vite ou diminue moins vite quand x augmente que la probabilité d'appartenance à la catégorie de référence. Une valeur positive ne signifie donc pas automatiquement que la probabilité d'appartenance à la catégorie j augmente quand x augmente, comme c'est le cas pour la régression binaire.

L'interprétation des valeurs numériques obtenues lors de la régression logistique doit bien entendu tenir compte de la catégorie choisie comme référence. Cependant, ce choix n'a pas fondamentalement d'importance, car à partir des résultats obtenus pour une catégorie de référence, on peut très simplement obtenir les résultats pour une autre catégorie de référence, sans refaire une nouvelle maximisation de la fonction de vraisemblance.

Si on souhaite exprimer les résultats en prenant par exemple comme référence la catégorie 1 alors qu'on dispose des résultats pour la catégorie de référence k, les nouvelles équations, notées g'_j s'écriront :

$$g'_j = g_j - g_1 \text{ et } g'_k = -g_1.$$

En effet :

$$\pi_j / \pi_1 = \exp(g'_j) = \frac{\pi_j / \pi_k}{\pi_1 / \pi_k} = \frac{\exp(g_j)}{\exp(g_1)} = \exp(g_j - g_1)$$

et $$\log_e (\pi_j / \pi_1) = g'_j = g_j - g_1.$$

Pour les données relatives au dépérissement, en prenant la classe 4 comme référence, on obtient les équations suivantes :

$$g_1 = 6{,}3514 - 0{,}016579\, x,$$
$$g_2 = 4{,}9582 - 0{,}008987\, x$$
et $$g_3 = 2{,}5449 - 0{,}004436\, x.$$

Le **tableau 2** donne les probabilités d'appartenance à chacune des catégories en fonction de l'altitude. Il donne également les *odds*, c'est-à-dire les rapports des probabilités π_j / π_4. On peut constater que, contrairement à la régression binaire, les probabilités d'appartenance à une classe ne sont plus obligatoirement toujours croissantes ou toujours décroissantes. Ainsi, la probabilité π_2 est d'abord croissante et ensuite décroissante. De même, la probabilité π_3 décroîtrait à partir d'une certaine altitude, si le domaine de variation de l'altitude était élargi. Par contre, la probabilité π_1 est toujours décroissante et la probabilité π_4 est toujours croissante avec l'altitude.

Les *odds ratios*, calculés pour une différence d'altitude de 100 m, valent respectivement 0,19, 0,40

Tableau 2. Modèle logit nominal : probabilités pour les quatre catégories de dépérissement (π_j) et *odds* (π_j/π_4) en fonction de l'altitude — *Nominal logit model : probability of the four decline categories* (π_j) *and odds* (π_j/π_4) *as a function of altitude*.

Altitude	π_1	π_2	π_3	π_4	π_1 / π_4	π_2 / π_4	π_3 / π_4
200	0,411	0,466	0,104	0,020	20,814	23,590	5,248
300	0,221	0,535	0,188	0,056	3,966	9,602	3,368
400	0,097	0,500	0,276	0,128	0,756	3,909	2,161
500	0,035	0,386	0,336	0,243	0,144	1,591	1,387
600	0,011	0,253	0,347	0,390	0,027	0,648	0,890

et 0,64. Ainsi, quand l'altitude augmente de 100 m, le rapport entre la probabilité d'être dans une classe et la probabilité d'être dans la classe 4 est divisé par 5 pour la classe 1, par 2,5 pour la classe 2 et par 1,6 pour la classe 3. On retrouve également ces valeurs en faisant le rapport de deux valeurs successives dans les trois dernières colonnes du **tableau 2**.

On notera que, dans le traitement de cet exemple, on n'a pas pris en compte le caractère ordinal de la variable *y* : on a considéré quatre classes sans tenir compte du caractère croissant du dépérissement quand on passe du niveau 1 au niveau 4. Les modèles décrits dans les trois paragraphes suivants vont, au contraire, intégrer cette information.

4. MODÈLE LOGIT BASÉ SUR LES PROBABILITÉS CUMULÉES

Lorsque la variable *y* traduit l'appartenance à l'une parmi *k* catégories ordonnées et qu'on souhaite tenir compte de l'ordre de ces catégories, différentes probabilités peuvent être modélisées.

Une première solution prend en compte les probabilités cumulées :

$$\omega_j = P\,(y \le j) = \pi_1 + ... + \pi_j,\ (j = 1, ..., k - 1).$$

Les logits s'écrivent :

$$g_j = \log_e [\omega_j / (1 - \omega_j)] = \log_e [(\pi_1 + ... + \pi_j) / (\pi_{j+1} + ... + \pi_k].$$

Ces logits sont appelés logits cumulés (*cumulative logits*) et, pour une catégorie donnée, ils correspondent aux logits utilisés dans la régression binaire, pour laquelle une catégorie correspondrait aux individus tels que $y \le j$ et l'autre catégorie correspondrait aux individus, tels que $y > j$.

Les logits sont alors exprimés en fonction de x par les relations : $g_j = \alpha_j + \beta\, x$, ajustées par la méthode du maximum de vraisemblance. On remarque que, dans ce modèle, chaque logit cumulé a sa propre ordonnée à l'origine, les α_j variant avec j. Par contre, le coefficient de régression β est le même pour toutes les relations. La prise en compte de coefficients de régression constants repose sur une hypothèse de parallélisme qui a l'avantage de conduire à une réduction du nombre de paramètres dans le modèle. La pertinence de cette hypothèse simplificatrice peut éventuellement être testée. Des informations à ce sujet sont données dans la discussion (§ 7).

Les *odds* sont égaux à :

$$(\pi_1 + ... + \pi_j) / (\pi_{j+1} + ... + \pi_k) = \exp(g_j)$$

et les *odds ratios* relatifs à x_i et x_{i+1} :

$$\frac{P(y \le j \,|x_i + 1) / [1 - P(y \le j \,|x_i + 1)]}{P(y \le j \,|x_i) / [1 - P(y \le j \,|x_i)]} = \exp(\beta)$$

sont indépendants, non seulement de la valeur de x, mais aussi de la catégorie considérée.

À partir des logits cumulés, on peut calculer les probabilités cumulées :

$$P(y \le j) = \pi_1 + ... + \pi_j = \exp(g_j) / [1 + \exp(g_j)].$$

Disposant des probabilités cumulées, il est évidemment possible d'obtenir, par différence, les probabilités relatives à une classe :

$$\pi_1 = P(y \le 1)$$
$$\pi_j = P(y \le j) - P(y \le j - 1),\ (j = 2, ..., k - 1)$$
et $$\pi_k = 1 - (\pi_1 + ... + \pi_{k-1}).$$

Pour les données de l'exemple, les équations obtenues sont les suivantes :

$$g_1 = 1{,}1837 - 0{,}008169\, x,$$
$$g_2 = 3{,}6514 - 0{,}008169\, x$$
et $$g_3 = 5{,}2426 - 0{,}008169\, x.$$

Le **tableau 3** donne les probabilités d'appartenance aux catégories et les *odds*, c'est-à-dire les rapports des probabilités : $P(y \le j) / [1 - P(y \le j)],\ (j = 1, ..., 3)$.

L'*odds ratio*, pour un accroissement de 100 m d'altitude, est égal à : $\exp[(100)(-0{,}008169)] = 0{,}442$, ce qui signifie que, d'une ligne à l'autre du **tableau 3,** les *odds* sont divisés par un facteur égal à 2,3 environ.

Tableau 3. Modèle logit basé sur les probabilités cumulées : probabilités pour les quatre catégories de dépérissement (π_j) et *odds* [exp(g_j)] en fonction de l'altitude — *Cumulative logit model: probabilities of the four decline categories* (π_j) *and odds [exp(g_j)] as a function of altitude*.

Altitude	π_1	π_2	π_3	π_4	exp (g_1)	exp (g_2)	exp (g_3)
200	0,389	0,493	0,091	0,026	0,638	7,521	36,92
300	0,220	0,549	0,174	0,058	0,282	3,323	16,31
400	0,111	0,484	0,283	0,122	0,124	1,468	7,207
500	0,052	0,341	0,368	0,239	0,055	0,649	3,184
600	0,024	0,199	0,362	0,416	0,024	0,287	1,407

La **figure 1** donne les probabilités cumulées en fonction de x. Le domaine de variation de x a été volontairement élargi au-delà des valeurs observées afin de mettre en évidence une caractéristique générale des courbes de probabilités cumulées : ces courbes ont la même forme et un déplacement horizontal permettrait de les faire coïncider. Cette situation résulte de l'utilisation dans le modèle d'un coefficient de régression β unique.

5. MODÈLE LOGIT BASÉ SUR LES CATÉGORIES ADJACENTES SUPÉRIEURES CUMULÉES

Si l'objectif est de comprendre ce qui distingue les individus qui ont atteint une catégorie donnée mais qui n'atteindront pas les catégories suivantes, on peut s'intéresser aux probabilités suivantes :

$$\omega_j = P(y = j | y \geq j) = \pi_j / (\pi_j + ... + \pi_k), (j = 1, ..., k - 1).$$

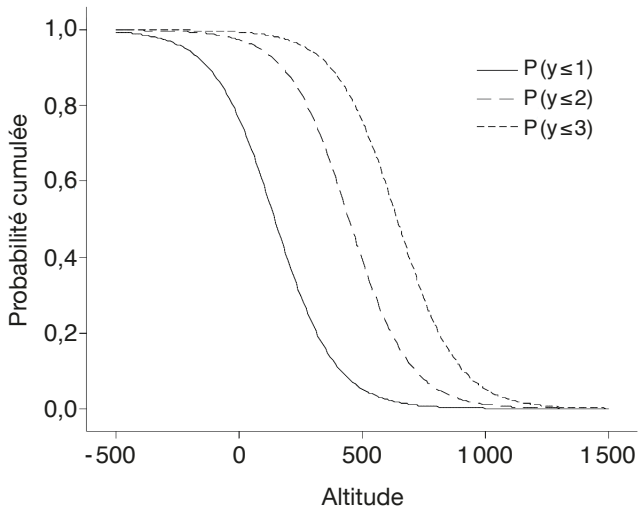

Figure 1. Modèle logit basé sur les probabilités cumulées : évolution des probabilités cumulées en fonction de l'altitude — *Cumulative logits model: cumulative probabilities as a function of altitude*.

Les logits (*continuation-ratio logits*) s'écrivent :

$$g_j = \log_e [\omega_j / (1 - \omega_j)] = \log_e [\pi_j / (\pi_{j+1} + ... + \pi_k)]$$

et correspondent aux logits utilisés dans la régression binaire pour laquelle une catégorie serait formée des individus pour lesquels $y = j$ et l'autre catégorie des individus pour lesquels $y > j$.

Ils sont exprimés en fonction de x par les relations : $g_j = \alpha_j + \beta x$, les coefficients de régression des différents logits étant considérés, par hypothèse, comme identiques.

Les *odds* sont donnés par les relations suivantes :

$$\pi_j / (\pi_{j+1} + ... + \pi_k) = \exp(g_j)$$

et le rapport entre les *odds* relatifs à x_i et à $x_i + 1$ sont indépendants de x et de la catégorie concernée :

$$\frac{P(y = j | x_i + 1) / P(y > j | x_i + 1)}{P(y = j | x_i) / P(y > j | x_i)} = \exp(\beta).$$

Les probabilités conditionnelles sont données par :

$$P(y = j | y \geq j) = \exp(g_j) / [1 + \exp(g_j)]$$

et les probabilités par catégorie sont obtenues par les relations suivantes :

$$\pi_1 = P(y = 1 | y \geq 1)$$
$$\pi_j = P(y = j | y \geq j) [1 - P(y < j)], (j = 2, ..., k - 1)$$
et $$\pi_k = 1 - (\pi_1 + ... + \pi_{k-1}).$$

Dans le cas où on ne souhaiterait pas imposer l'égalité des coefficients de régression des relations donnant les logits, on peut montrer que l'estimation au sens du maximum de vraisemblance de l'ensemble des paramètres est identique à l'estimation qui serait obtenue par la réalisation de $k - 1$, ajustements indépendants sur les données binaires $y = j$ et $y > j$, comme expliqué ci-dessus. Cette équivalence est d'ailleurs utilisée en pratique pour l'ajustement du modèle à l'aide des logiciels statistiques, y compris

lorsqu'on souhaite un coefficient de régression unique, l'ajustement se faisant alors après un recodage des données.

Pour l'exemple, on obtient les logits suivants :

$$g_1 = 0{,}7595 - 0{,}006730\,x,$$
$$g_2 = 2{,}8542 - 0{,}006730\,x$$
et $$g_3 = 3{,}4897 - 0{,}006730\,x.$$

Le **tableau 4** donne les probabilités d'appartenance aux différentes catégories ainsi que les *odds*, en fonction de l'altitude.

Pour un accroissement de 100 m d'altitude, l'*odds ratio* vaut : exp[(100)(-0,00673)] = 0,510.

Le rapport entre la probabilité qu'un chêne soit dans la catégorie j et la probabilité qu'il soit dans une catégorie supérieure à j est divisé par deux quand l'altitude augmente de 100 m, comme le montrent les trois dernières colonnes du **tableau 4**.

Si on calculait les probabilités conditionnelles :

$$P(y = j \,|\, y \geq j),\ (j = 1, ..., k - 1),$$

et qu'on les portait sur un graphique en fonction de x, on obtiendrait trois sigmoïdes décroissantes qui, comme dans le cas de la **figure 1**, peuvent être superposées par un déplacement horizontal. Ces sigmoïdes seraient cependant légèrement moins redressées que dans la **figure 1**.

6. MODÈLE BASÉ SUR LES LOGITS DES CATÉGORIES ADJACENTES

Une troisième approche pour les données ordinales prend en compte les probabilités suivantes :

$$\omega_j = P(y = j | y = j \text{ ou } j + 1) = \pi_j / (\pi_j + \pi_{j+1}),$$
$$(j = 1, ..., k - 1),$$

pour lesquelles les logits sont :

$$g_j = \log_e [\omega_j / (1 - \omega_j)] = \log_e (\pi_j / \pi_{j+1}).$$

Ces logits, appelés logits des catégories adjacentes (*adjacent-categories logits*), sont équivalents aux logits utilisés dans la régression binaire pour laquelle la catégorie correspondant à l'évènement serait la catégorie $y = j$ et l'autre catégorie serait la catégorie $y = j + 1$. Ils sont exprimés en fonction de x, en considérant, ici aussi, que les coefficients de régression sont constants : $g_j = \alpha_j + \beta\, x$, $(j = 1, ..., k - 1)$.

Bien que le modèle des catégories adjacentes soit un cas spécial du modèle logit nominal, celui-ci ne permet pas d'imposer les contraintes appropriées par la méthode du maximum de vraisemblance. Par contre, si les données sont groupées dans une table de contingence, il peut être ajusté par la méthode des moindres carrés pondérés (Allison, 1999).

Les *odds* sont égaux à : $\pi_j / \pi_{j+1} = \exp(g_j)$ et les *odds ratios* sont identiques pour les différentes catégories de y :

$$\frac{\pi_j (x_i + 1) / \pi_{j+1} (x_i + 1)}{\pi_j (x_i) / \pi_{j+1} (x_i)} = \exp(\beta).$$

À partir des *odds*, c'est-à-dire des rapports de probabilités pour des catégories adjacentes :

$$\pi_j / \pi_{j+1} = \exp(g_j),$$

on peut retrouver les probabilités d'appartenance aux catégories : $\pi_j = \exp(g_j + ... + g_{k-1}) / D$

avec :

$$D = 1 + [\exp(g_1 + ... + g_{k-1}) + \exp(g_2 + ... + g_{k-1}) + ... + \exp(g_{k-1})]$$
et $\pi_k = 1 - (\pi_1 + ... + \pi_{k-1})$.

Pour ajuster le modèle aux données de l'exemple avec le logiciel SAS, les altitudes ont été regroupées en trois classes : la classe 1 comprend les valeurs inférieures à 285 m, la classe 2 les valeurs comprises entre 285 m et 400 m et la classe 3, les valeurs supérieures à 400 m. Les trois classes ainsi obtenues

Tableau 4. Modèle logit basé sur les catégories supérieures cumulées : probabilités pour les quatre catégories de dépérissement (π_j) et *odds* [exp(g_j)] en fonction de l'altitude — *Continuation-ratio logits model: probabilities of the four decline categories* (π_j) *and odds [exp(g_j)] as a function of altitude*.

Altitude	π_1	π_2	π_3	π_4	exp (g_1)	exp (g_2)	exp (g_3)
200	0,357	0,526	0,104	0,012	0,556	4,518	8,534
300	0,221	0,543	0,192	0,044	0,284	2,305	4,354
400	0,126	0,472	0,277	0,125	0,145	1,176	2,221
500	0,069	0,349	0,309	0,273	0,074	0,600	1,133
600	0,036	0,226	0,270	0,468	0,038	0,306	0,578

présentent des effectifs de même ordre de grandeur.

Les résultats suivants ont été obtenus :

$g_1 = 0{,}2795 - 0{,}6633\ x,$
$g_2 = 2{,}2027 - 0{,}6633\ x$
et $g_3 = 2{,}2950 - 0{,}6633\ x.$

Le **tableau 5** donne les probabilités d'appartenance aux différentes catégories ainsi que les *odds*, c'est-à-dire les rapports des probabilités des classes successives.

L'*odds ratio* est égal à : exp(-0,6633) = 0,515.

Cela signifie donc que le rapport entre la probabilité d'être dans une catégorie et la probabilité d'être dans la catégorie immédiatement supérieure est divisée par deux quand on augmente d'une classe d'altitude, comme le montrent les résultats du **tableau 5**.

7. DISCUSSION

Pour une variable à expliquer *y* à plus de deux catégories ordonnées, nous avons présenté trois modèles (§ 4, 5 et 6). Ceux-ci diffèrent par les probabilités et donc aussi par les logits qui sont modélisés. Le **tableau 6** résume les situations examinées. Il reprend également le modèle nominal, à titre de comparaison.

Il faut tout d'abord noter que d'autres modèles équivalents à ceux présentés auraient pu être définis. Ainsi, par exemple, pour le modèle basé sur les logits cumulés, on pourrait s'intéresser aux probabilités $P\ (y > j)$ pour lesquelles les logits seraient :

$$\log_e\ [(\pi_{j+1} + ... + \pi_k)] / (\pi_1 + ... + \pi_j)].$$

Par rapport au cas repris au § 4, le signe des coefficients des équations donnant les logits serait changé. De même, dans le cas des catégories adjacentes, on pourrait définir les probabilités d'intérêt :

$$P(y = j + 1 | y = j \text{ ou } j + 1),$$

ce qui conduirait également à un changement de signe des coefficients des fonctions logistiques, par rapport au modèle présenté au § 6.

D'autre part, les modèles examinés permettent tous d'estimer les probabilités d'appartenance aux différentes catégories π_j. Pour l'exemple traité, on constate d'ailleurs une forte similarité des π_j pour le modèle nominal et pour les deux premiers modèles pour catégories ordinales, le troisième modèle ne pouvant pas être comparé car la variable altitude a été groupée en trois classes.

Pour les différents modèles proposés, on a toujours considéré, par simplicité, que les logits évoluent de façon linéaire en fonction d'une seule variable *x*. Comme signalé dans l'introduction, les modèles peuvent être étendus au cas de plusieurs prédicteurs. Ces prédicteurs peuvent correspondre à des caractéristiques différentes ou à des fonctions de caractéristiques, comme des variables élevées au carré pour tenir compte d'effets non linéaires ou des produits de variables pour tenir compte d'interactions.

Comparée au modèle pour données nominales, la prise en compte de l'ordre des catégories permet de réduire le nombre de paramètres, puisqu'on considère que les coefficients de régression sont constants pour les différentes fonctions logistiques. Cette parcimonie dans le nombre de paramètres est un avantage pour

Tableau 5. Modèle basé sur les logits des catégories adjacentes : probabilités pour les quatre catégories de dépérissement (π_j) et *odds* (π_j / π_{j+1}) pour les trois classes d'altitude — *Adjacent-categories logit model: probabilities of the four decline categories* (π_j) *and odds* (π_j / π_{j+1}) *as a function of altitude*.

Classes d'altitude	π_1	π_2	π_3	π_4	π_1 / π_2	π_2 / π_3	π_3 / π_4
1	0,352	0,516	0,111	0,022	0,681	4,662	5,113
2	0,182	0,519	0,216	0,082	0,351	2,402	2,634
3	0,070	0,387	0,313	0,230	0,181	1,237	1,357

Tableau 6. Probabilités modélisées dans les différents modèles et *odds* correspondants — *Probabilities modelized by the different models and related odds*.

Modèles	**Probabilités**	***Odds***
Nominal	$P(y = j \mid y = j \text{ ou } k)$	π_j / π_k
Logits cumulés	$P\ (y \leq j)$	$(\pi_1 + ... + \pi_j) / (\pi_{j+1} + ... + \pi_k)$
Catégories adjacentes supérieures cumulées	$P(y = j \mid y \geq j)$	$\pi_j / (\pi_{j+1} + ... + \pi_k)$
Logits des catégories adjacentes	$P(y = j \mid y = j \text{ ou } j + 1)$	π_j / π_{j+1}

autant que l'hypothèse de parallélisme des fonctions logistiques soit réaliste. Ce parallélisme des fonctions correspond à la constance des *odds ratios*. Dans les **tableaux 3**, **4** et **5**, la constance des *odds ratios* s'est traduite par un rapport constant entre valeurs successives dans les trois dernières colonnes de chacun des tableaux, quelle que soit la colonne considérée, et la différence entre les trois modèles se situe dans la définition des *odds* (**tableau 6**).

On notera qu'il existe des tests permettant de vérifier les hypothèses de parallélisme. Des informations à ce sujet et sur la stratégie de recherche de modèles lorsque l'hypothèse de parallélisme est rejetée sont données par Brant (1990), Clogg et al. (1994), Long (1997), Allison (1999) et O'Connell (2000).

Indépendamment du respect du parallélisme des fonctions logistiques, le choix d'un modèle de régression ordinale préférentiellement à un autre dépend aussi de la simplicité de l'interprétation des résultats, en relation avec l'objectif de l'étude. Si l'interprétation d'un type d'*odds ratios* particulier est un objectif majeur de l'étude, c'est ce type d'*odds ratios* qui sera évidemment retenu pour la simplicité qui en découle dans l'interprétation. En effet, même s'il est toujours possible de calculer *a posteriori* des *odds ratios* d'un type donné à partir des π_j, alors qu'un autre type a été utilisé dans les modèles, il est évidemment beaucoup plus simple d'interpréter une seule valeur que d'être confronté à des *odds ratios* qui varient avec x et avec j.

Ainsi, on préférera le modèle basé sur les logits cumulés si le but de l'étude est de clarifier la tendance, plutôt croissante ou décroissante, en fonction de x, de la probabilité d'avoir atteint un niveau donné de y (Agresti, 2002). Le modèle basé sur les catégories adjacentes supérieures cumulées sera plus utile dans des études de développement lorsque l'objectif principal est l'identification des facteurs associés au fait d'être dans une catégorie faible, sachant qu'un certain niveau a déjà été atteint (O'Connell, 2000). Les logits des catégories adjacentes seront utiles pour rechercher quelles sont les variables explicatives qui peuvent au mieux prédire la probabilité d'être dans une catégorie plutôt que dans la catégorie suivante.

La réalisation des calculs ne pose pas de problèmes particuliers avec les logiciels tels que SAS, Minitab ou R pour la régression binaire, la régression nominale et la régression ordinale basée sur les logits cumulés. Pour ces modèles, les procédures ou les commandes des logiciels donnent directement les résultats. Pour le modèle basé sur les logits des catégories supérieures cumulées, les logiciels nécessitent un recodage des données, suivi de l'ajustement d'un modèle binaire sur données recodées.

Quant au modèle basé sur les logits des catégories adjacentes, il n'est pas disponible dans Minitab. Par contre, il peut être ajusté par la procédure CATMOD de SAS ou par la fonction ACAT de R, qui utilisent la méthode des moindres carrés pondérés dont le point de départ est une table de contingence (Clogg et al., 1994 ; SAS Institute, 1995 ; Allison, 1999 ; Stokes et al., 2000). C'est la raison pour laquelle, au § 6, la variable altitude a été recodée.

Enfin, on notera également que les options par défaut des logiciels peuvent varier et conduire à des résultats à première vue différents mais en réalité tout à fait équivalents. Ces variantes peuvent se marquer, par exemple, sur les modalités de référence ou sur la définition des probabilités qui sont modélisées.

Bibliographie

Agresti A., 2002. *An introduction to categorical data analysis*. New York, USA: Wiley.

Allison P.D., 1999. *Logistic regression using SAS system: theory and application*. Cary, NC, USA: SAS Institute.

Brant R., 1990. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, **46**, 1171-1178.

Clogg C.C. & Shihadeh E.S., 1994. *Statistical models for ordinal variables*. Thousand Oaks, CA, USA: Sage.

Collett D., 1999. *Modelling binary data*. London: Chapman & Hall/CRC.

Duyme F. & Claustriaux J.J., 2006. *La régression logistique binaire. Notes de statistique et d'informatique*. Gembloux, Belgique : Faculté universitaire des Sciences agronomiques de Gembloux.

Gillet A., 2005. *Influences stationnelle, sylvicole et spécifique sur le dépérissement des chênes indigènes (*Quercus robur *L. et* Quercus petraea *[Matt.] Liebl.) en Région Wallonne*. Mémoire : Faculté universitaire des Sciences agronomiques de Gembloux (Belgique).

Gillet A., 2007. *Régression logistique polychotomique ordinale*. Travail de fin d'étude réalisé dans le cadre du diplôme d'Études approfondies en Statistique et Informatique appliquées : Faculté universitaire des Sciences agronomiques de Gembloux (Belgique).

Hosmer D.W. & Lemeshow S., 2000. *Applied logistic regression*. New York, USA: Wiley.

Long J.S., 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA, USA: Sage.

O'Connell A., 2000. Methods for modelling ordinal outcome variables. *Meas. Eval. Counseling Dev.*, **33**(3), 170-193.

SAS Institute, 1995. *Logistic regression examples using SAS system*. Cary, NC, USA: SAS Institute.

Stokes M.E., Davsi C.S. & Koch G.G., 2000. *Categorical analysis using SAS System*. 2nd ed. Cary, NC, USA: SAS Institute.

(13 réf.)