INTERPRETATION DES MODELES LOGISTS MULTINOMIALS ORDONNEES EN TERMES DE DISTRIBUTIONS.

Malika CHIKHI[1], Thierry MOREAU[2], Michel CHAVANCE[2].

1. Faculté des Sciences Exactes, Département de Mathématiques, Université Constantine (Algérie).
2. INSERM, unité 472, Avenue P. Vaillant Couturier, Villejuif (France).

Reçu le 17/07/2008 – Accepté le 28/11/2011

Résumé

Cet article met au point l'interprétation en termes de distribution de deux modèles logistiques : le modèle cumulatif et le modèle à rapports continus pour une variable réponse ordinale à c catégories.

Ces modèles seront appliqués aux données d'une étude épidémiologique portant sur la consommation d'alcool que nous analyserons et interpréterons.

Mots clés: statistique, modèle logistique, logit cumulatif, logit à rapport continu.

Abstract

This paper reviews interpretation of two continuation-ratio logits models for ordinal response variables to C categories in term of distribution.

These models will be applied to data of an epidemiological survey of alcohol consumption that we will analyze and will interpret.

Keywords: *statistical, logit model, cumulative logit, continuatio-ratio logit.*.

ملخص

هذا المقال يتناول تطور تفسير من حيث توزيع نموذجين لوجستية: نموذج المضافة ونموذج لعلاقة مستمرة مع ترتيبي استجابة متغيردات C فئات.

وسيتم تطبيق هذه النماذج على بيانات من دراسة وبائية على استهلاك الكحول بتوضيح تحليلها وتفسيرها.

الكلمات المفتاحية : الاحصاء ، logit التراكمي، logit مستمر

Introduction

Bien que décrite depuis plusieurs années, l'extension de la régression logistique au cas où la variable expliquée est catégorielle ordonnée n'est pas toujours utilisée quand elle pourrait l'être. L'objectif de cette note est de rappeler les différents modèles convenant à cette estimation ; ils sont décrits en détail dans l'ouvrage de référence de A. Agresti (1,2).

Un aspect de certains de ces modèles, qui peut être utile pour leur mise en application, est particulièrement souligné ; il s'agit de leur interprétation dans le cas où les classes de la variable expliquée peuvent être considérées comme provenant de la partition de l'intervalle de variation d'une variable aléatoire sous-jacente continue.

Pour plusieurs modèles, l'interprétation est rappelée et détaillée, pour l'un d'entre eux la famille de distribution correspondante est déduite.

Une étude épidémiologique portant sur la consommation d'alcool chez les adolescents présentée par ailleurs (2) et dont les résultats seront partiellement réanalysés et commentés. Pour appliquer certaines des méthodes évoquées, les méthodes utilisées sont celles qui sont disponibles dans le logiciel BMDP (4).

MODELE LOGISTIQUE GENERAL POUR UNE VARIABLE REPONSE A c CATEGORIES ORDONNEES.

Présentation d'un exemple provenant d'une étude épidémiologique.

L'étude condensée avait pour but d'identifier et d'analyser les facteurs de risque de consommation d'alcool chez les adolescents. Les renseignements, recueillis par questionnaire, ont conduit à considérer quatre classes ordonnées de consommation prenant les valeurs suivantes : non buveur, petit buveur, moyen buveur et grand buveur ; les facteurs de risque concernent les caractéristiques individuelles (sexe…) et les caractéristiques de la vie sociale (statut matrimonial des parents …).

Il s'agit d'évaluer l'effet de chaque facteur sur la probabilité que la variable « alcool » prenne une valeur ou un ensemble de valeurs plutôt qu'une autre valeur ou un autre ensemble de valeurs rapportées à la probabilité que cette variable prenne une autre valeur ou un autre ensemble de valeurs.

Soit x' $(x_1, \dots, x_k)$ un vecteur de k variables explicatives et y la variable expliquée pour laquelle c classes ordonnées sont définies et supposées être en réalité continues ; les différents logits ci-dessous exprimés ici conditionnellement à x s'écrivent :

$$L_j(x) = \log \frac{pr(y \geq j+1/x)}{pr(y \leq j/x)} \quad : \quad \text{logit cumulatif} \qquad (1)$$

$$L_j(x) = \log \frac{pr(y = j+1/x)}{pr(y \leq j/x)} \quad : \quad \text{logit à rapport continu} \qquad (2)$$

$$L_j(x) = \log \frac{pr(y = j+1/x)}{pr(y = j/x)} \quad : \quad \text{logit à catégories adjacentes} \qquad (3)$$

On remarque que ces trois logits sont identiques dans le cas où c =2. Le modèle général est donné par :

$$L_j(x) = \alpha_{j+} \beta' x$$

β' est un vecteur de paramètres inconnus ($\beta_1, \dots, \beta_k$) à estimer ainsi que les α_j avec

j = 1, … , c-1.

Dans le cas où l'une des variables explicatives (par exemple x_1) est catégorielle ordinale et si un score u_i peut être affecté à chaque catégorie i avec i = 1,…, r de x_1 on peut écrire :

$$Lj\,(x=u_i\,,\,x_2\,,\,\dots\,,\,x_k)\;\; Lj\,(x=u_{i-1}\,,\,x_2\,,\,\dots\,,\,x_k) = \beta(u_i - u_{i-1}).$$

Il apparaît ainsi que les $\exp\beta(u_i - u_{i-1})$ s'interprètent comme des odds-ratio et le modèle suppose que ces odds-ratio sont identiques pour tout j ; en particulier dans le cas où les scores u_i sont entiers consécutifs $u_i = i$, l'odds-ratio vaut exp(β) ; si $L_j(x)$ est donnée par (1), c' est l'odds-ratio entre la variable binaire définie par :

$F_j(x) = pr(y \leq j/x)$ et $1 - F_j(x)$;

et la variable binaire définie par l'appartenance à l'une ou l'autre de deux catégories adjacentes de x_1.

- Dans le cas où x_1 est la seule variable explicative ordinale comme ci-dessus, le modèle a été dénommé par Agresti (1) : *modèle logistique d'association uniforme*.
- Dans le cas où l'une des variables explicatives est qualitative nominale à r catégories, le modèle générale s'applique en définissant par exemple (r-1) variables binaires indicatrices de chaque catégories, en choisissant une catégorie de référence pour laquelle les (r-1) variables sont nulles. Il faut noter que lorsqu'une seule variable explicative de ce type est considérée, le modèle se dénomme : *modèle à effets lignes* (5);

- les modèles utilisant les logits à catégories adjacentes ont surtout été présentés par Goodeman (6).

Parmi les auteurs qui ont étudié les modèles sous différentes formes, on peut citer Snell (7), Williams et Grizzle (8), Mc Cullagh (9) et Bock (10).

$$F_x(a_j) = \frac{1}{1+\exp(\alpha_{j+} \beta x)}$$

Les auteurs qui ont suggéré le modèle logit à ratio continu sont Thompson (11), Cox (12), Fienberg et Masson (13) et Mc Cullagh et Nelder (14).

ESTIMATIONS DES PARAMETRES DU MODELE.

Les auteurs qui ont suggéré le modèle logistique à effets lignes sont Williams et Grizzle (8). Ils ont ajusté ces modèles en utilisant la méthode des moindres carrés pondérés.

En 1973, Bock et Yates ont utilisé la méthode du maximum de vraisemblance (15).

TEST DE SIGNIFICATION DES PARAMETRES.

L'hypothèse portant sur la nullité d'un ou plusieurs paramètres du modèle logistique général peut être testée par un test du log du rapport des vraisemblances.

Dans le cas des variables explicatives catégorielles, le test est analogue au test du chi-deux défini précédemment. Il s'interprète de la même manière en

$$F_x(a_j) = \frac{1}{1+\exp(-a_{j+} \beta x)}$$

termes d'indépendance conditionnelle entre les variables explicatives et la variable réponse.

INTERPRETATION DES MODELES LOGISTIQUES AVEC VARIABLE REPONSE ORDINALE.

Dans ce paragraphe, les hypothèses du modèle logistique,[formules (1), (2) et (3)], sont explicitées en termes de distribution de la variable réponse Y conditionnelle aux variables explicatives. Deux modèles particuliers seront considérés correspondant respectivement

$$F_x(a_j) = \text{pr}(Y \le a_j / X = x)$$

au logit cumulatif et au logit à rapport continu.

DEFINITIONS, NOTATIONS ET RAPPELS.

La variable expliquée Y, pour laquelle c classes ordonnées sont définies, est supposée être en réalité continue, les différentes classes constituant une partition de son intervalle de variation. Cette situation correspond à la plupart des cas rencontrés en pratique.

En notant $a_1 < a_2 < \ldots < a_c$, les limites (observées ou non) des c intervalles contigus où Y prend ses valeurs, la probabilité que Y appartienne à la classe j sera :

$\pi_j = \text{pr}(a_{j-1} < Y < a_j)$,

et la fonction de répartition de Y en a_j (j = 1,…, c) s'écrit :

$F(a_j) = \text{pr}(Y \le a_j) = \pi_1 + \ldots + \pi_j$.

$$\frac{1-F_x(a_j)}{F_x(a_j)} = \exp(\alpha_{j+} \beta x) \qquad (4)$$

Soit X une variable explicative quantitative ou qualitative. Les calculs sont effectués ci-dessous avec la seule variable explicative X mais seraient inchangés si d'autres variables étaient incluses dans le modèle.

et $S_x(a_j) = 1 - F_x(a_j)$ est la fonction de survie correspondante.

La fonction de répartition de Y conditionnelle à X en $Y = a_j$ est :

LE MODELE A LOGITS CUMULE.

Le modèle s'écrit (1) :

$$\frac{Sx_2(a_j)/F_{x2}(a_j)}{Sx_1(a_j)/F_{x1}(a_j)} = \exp \beta(x_2 - x_1) \quad ; \quad j = 1,\ldots,c. \qquad (5)$$

avec :

Soient x_1 et x_2 deux valeurs prises par X et telles que $x_1 < x_2$, d'après le modèle :

$$Fx_1(a_j) > Fx_2(a_j) \ , \ j = 1,\ldots,c \, .$$

Si $\beta \ge 0$, il s'ensuit que le rapport ci-dessus est supérieur à 1 pour tout j ce qui entraîne :

La distribution de Y en chaque a_j , conditionnelle à x_2 apparaît alors stochastique ment supérieure à celle de Y conditionnelle à x_1.

Si $\beta < 0$, c'est l'inverse qui est vrai .

Plus précisément l'expression (5) définit une relation entre les distributions de Y conditionnelles à x_1 et x_2.

$$Fx_2(a_j) = Fx_1\left[a_{j+} \beta(x_2 - x_1)\right].$$

Cette relation est en particulier vérifiée pour des distributions logistiques translatées l'une de l'autre. En effet si la distribution de Y conditionnelle à x est supposée logistique, on peut toujours écrire :

Le modèle est alors vérifié en posant $\alpha_j = - a_j$ et le paramètre de translation entre les distributions de Y conditionnelles à x_1 et à x_2 est $\exp(\beta x_2 - x_1)$. En effet, il apparaît que :

On peut remarquer que dans le cas où x est une variable ordinale prenant des valeurs entières consécutives, $\exp(\beta)$ est l'odds-ratio « local global » défini précédemment entre deux variables consécutives de x et pour tout un j donné.

LE MODELE A LOGIT A RAPPORT CONTINU.

Dans le paragraphe précédent, la démarche a consisté à chercher la distribution de la variable aléatoire continue sous-jacente Y qui coïncide avec la fonction de répartition, définie par le modèle logit cumulé, aux points a_j qui sont les extrémités des classes définies par Y ; dans ce qui suit la démarche sera différente et consistera a trouver la limite continue du modèle discret défini pour les variables catégorielles. Deux modèles différents à logit à rapport continu peuvent être envisagés.

En notant :

$$Fx_1(a_j) = (\frac{a_j}{k})^{t \exp \beta x1},\ Fx_2(a_j) = (\frac{a_j}{k})^{t \exp \beta x2}$$

Le model est vérifié

$$f(y) = \frac{t}{k^t} y^{t-1}\ ;\ F(y) = \frac{y^t}{k^t} = (\frac{y}{k})t$$

$$p_j = \frac{pr(a_j < Y \le a_{j+1})}{S(a_j)} \longrightarrow \frac{f(a_j)}{S(a_j)} dy = \lambda(a_j)dy$$

$P_j = pr\ (a_j < Y \le a_{j+1}) / Y > a_j)$

et $P'_j = pr\ (a_j < Y \le a_{j+1}) / Y < a_{j+1})$

ils expriment le logit de p_j ou de p'_j conditionnels à X=x sous la forme : $(\alpha j + \beta x)$.

L'interprétation présentée concerne la limite du modèle lorsque a_{j+1} tend vers a_j pour tout j (et c tend vers l'infini) c'est-à-dire lorsque ce sont les réalisations de la variable continue Y qui sont observées.

Dans ce cas :

où $\lambda(a_j)$ est par définition la fonction de « risque instantané » de Y en a_j ; de même $(1-p_j)$ tend vers 1.

$$p'_j = \frac{pr(a_j < Y \le a_{j+1})}{F(a_{j+1})} \quad \text{tend vers} \quad \frac{f(a_j)}{F(a_j)} dy$$

De la même façon :

où $f(a_j)$ est la densité de probabilité de Y en a_j. De plus $1-p'_j$ tend vers 1.

Soient x_1 et x_2 deux valeurs prises par X et telles que $x_1 < x_2$, en conditionnant par rapport à X , les deux modèles s'écrivent respectivement :

$$\frac{fx(a_j)}{Sx(a_j)} = \exp(\alpha_j + \beta x) \qquad \text{et}$$

En écrivant le premier modèle pour x_1 et pour x_2,

$$\frac{fx_2(a_j)}{Sx_2(a_j)} dy = \frac{fx_1(a_j)}{Sx_1(a_j)} dy.\exp\beta(x_2 - x_1)$$

$$Fx_2(a_j) = \{Fx_1(a_j)\}^{\exp\beta(x2-x1)}$$

on obtient :

Soit après intégration :

$$Sx_2(a_j) = \{Sx_1(a_j)\}^{\exp\beta(x2-x1)}$$

Le deuxième modèle conduirait de même à :

Chacune de ces deux expressions définit une relation particulière entre les distributions de Y conditionnelles à x_1 et x_2.

Le premier modèle est connu sous le nom de « modèle de risques instantanés proportionnels » ou « modèle de Cox » (12) et il est couramment utilisé dans les études de survie ; un cas particulier de ce modèle est celui où : $S_x(y) = \exp\{(-\lambda y \exp(\beta x)\}$ correspondant à une distribution exponentielle pour la distribution de Y conditionnelle à X.

Nous avons pu trouver une famille de distribution satisfaisant le deuxième modèle : cette famille comprend la distribution de Pareto dont la densité et la fonction de répartition sont les suivantes :

En effet, en écrivant :

En conclusion, pour choisir entre ces différents modèles, on pourra tenir compte de ces résultats selon les informations à priori dont on dispose sur la distribution de la variable Y et la façon dont elle varie selon les catégories de la variable explicative.

Selon les informations de l'étude qu'on possède on peut choisir le premier modèle (logit cumulés) ; si, entre les différentes catégories de la variable X, la variable Y subit une translation, le deuxième ou le troisième modèle (à rapport continu) dans le cas de changement d'échelle.

APPLICATION EPIDEMIOLOGIQUE.

Introduction.

Plusieurs études épidémiologiques sont décrites et commentées quotidiennement dans la littérature (16,22), dans le même ordre d'idées notre étude avait pour but d'identifier et d'analyser les facteurs de risques de consommation d'alcool chez les adolescents (3), dont une

$$\frac{fx(a_j)}{Fx(a_j)} = \exp(\alpha_j + \beta x)$$

partie des données est réanalysée ci-dessous.

Les renseignements, recueillis par questionnaire, ont conduit à considérer quatre classes ordonnées de consommation ; les facteurs de risque concernent des caractéristiques individuelles (sexe…) et des caractéristiques de la vie sociale (statut matrimonial des parents …).

Les modèles explicatifs appliqués à ces consommations sont les modèles à logit cumulés et à catégories adjacentes.

L'intérêt de la comparaison de ces modèles provient du fait qu'ils sont disponibles dans le logiciel BMDP (4) et largement utilisables en pratique.

MATERIELS ET METHODES.

Matériel.

L'enquête a été effectuée en milieu scolaire (établissements secondaires publiques ou privés) dans deux villes de taille moyenne.

L'échantillon initial a été obtenu par tirage au sort à l'intérieur de chaque établissement. Chaque élève a rempli un auto-questionnaire portant en particulier sur sa situation sociodémographique, la relation familiale, son intégration sociale, sa santé et ses consommations d'alcool.

L'échantillon considéré ici comprend 417 adolescents ; à partir des renseignements contenus dans le questionnaire, cinq variables ont été constituées.

- *la variable consommation d'alcool* définie à partir des réponses à plusieurs questions concernant la

consommation plus ou moins régulière de différentes boissons ; cette variable comprend quatre classes : non buveurs notée (nb) par la suite, petits buveurs notée (pb), moyens buveurs notée (mb) et grands buveurs notée (gb), codées 0, 1, 2 et 3 respectivement.

- *La variable statut matrimonial des parents* notée « matripa » par la suite codée en 1 (parents mariés ou concubins) ou 0 (parents séparés, décédés, inconnus).
- *La variable indiquant la sociabilité* notée « social » par la suite a été constituée par la somme de six variables codées 0 ou 1 concernant par les sorties avec d'autres jeunes, la pratique de jeux d'argent ; cette variable comprend donc sept catégories codées de 0 à 6.
- *La variable indiquant la depressivité* notée « depress » par la suite est la somme de sept variables codées en 0 et 1 concernant par exemple l'ennui, les perturbations de sommeil ; cette variable comprend donc huit catégories codées de 0 à 7.

Par ailleurs, le sexe a été codé en 0 (masculin) et 1 (féminin) ; enfin les sujets appartenaient tous à la même classe d'âge (16-18 ans) et les âges individuels n'ont donc pas été pris en compte.

Méthodes.

Les modèles logistiques à logit cumulés et à catégories adjacentes ont été appliqués avec comme variables expliquées l'alcool et variables explicatives : sexe, matripa, dépress et social.

Le programme utilisé pour traiter ces données est le programme PR du logiciel BMDP (4).

RESULTATS ET COMMENTAIRES.

Description de l'échantillon.

Parmi les 417 sujets examinés, 229 sont des garçons et 188 des filles. Concernant la consommation d'alcool, le nombre des non buveurs est 81, de petits buveurs 214, de moyens buveurs 82 et de grands buveurs 40.

Facteurs de variation de consommation d'alcool.

Le tableau I montre les relations entre la variable alcool et les variables explicatives (sexe, matripa, dépress, social) pour le modèle logit à catégories adjacentes.

TABLEAU I. Odds-ratio entre la consommation d'alcool et les variables explicatives (modèle logit à catégories adjacentes) .

Variable	Odds-ratio	Intervalle de confiance à 95%	Degré de signification
Sexe	**1,90**	**(1,40 – 2,50)**	**$P < 10^{-5}$**
Matripa	**1,6**	**(1,10 – 2,30)**	**P < 0,019**
Depress	**0,82**	**(0,75 – 0,89)**	**$P < 10^{-5}$**
Social	**0,73**	**(0,67 – 0,80)**	**$P < 10^{-9}$**

Il apparaît que les effets de la dépressivité et du sexe sont équivalents et significatifs.

Le statut matrimonial des parents présente un effet significatif à 2%.

La sociabilité a un effet beaucoup plus significatif que les autres variables.

Plus précisément la « chance » d'être (pb/nb) ou (mb/pb) ou (gb/mb) chez les filles est 1,9 fois plus grande que chez les garçons.

La « chance » d'être (pb/nb) ou (mb/pb) ou (gb/mb) chez les enfants dont la famille est désunie est 1,6 fois plus grande que chez les enfants dont les parents sont unis.

Pour les mêmes catégories de buveurs comparées, le facteur multiplicatif est 0,82 lorsque la variable « dépressivité » augmente d'une unité et de 0,73 lorsque la variable « sociabilité » augmente d'une unité.

Le log de vraisemblance de ce modèle est : $L'_1 = -455{,}862$.

La statistique du test du log de rapport de vraisemblance est : $G'_1 = 279{,}02$ avec un nombre de degré de liberté $ddl'_1 = 308$ (degré de signification $p'_1 = 0{,}881$).

Le tableau II montre les relations entre les variables explicatives (sexe, matripa, dépress et social) et la variable alcool pour le modèle à logit cumulatif.

TABLEAU II. Odds-ratio entre la consommation d'alcool et les variables explicatives (modèle à logit cumulatif).

Variable	Odds-ratio	Intervalle de confiance à 95%	Degré de signification
Sexe	**0,40**	**(0,25 – 0,60)**	**$P < 10^{-5}$**
Matripa	**0,54**	**(0,31 – 0,92)**	**P < 0,021**
Depress	**1,30**	**(1,20 – 1,50)**	**$P < 10^{-5}$**
Social	**1,50**	**(1,40 – 1,80)**	**$P < 10^{-9}$**

Les significations associées aux variables explicatives sont du même ordre de grandeur que celles obtenues avec le modèle à catégories adjacentes exposées au tableau I.

Les odds-ratio s'interprètent de la façon suivante : la « chance » d'être (gb/nb, pb, mb) ou (gb, nb/pb, mb) ou (gb, mb, pb/nb) chez les filles est 0,40 fois plus faible que chez les garcons.

De même cette « chance » est 0,54 fois plus petite lorsque les parents sont désunis plutôt qu'unis.

Lorsque les variables « dépressivité » et « sociabilité » augmentent d'une unité, les chances comparées de ces catégories de buveurs sont multipliées par 1,3 et 1,5 respectivement.

Le log de vraisemblance est : $L'_2 = -455{,}41$.

La statistique du test du log du rapport de vraisemblance est : $G'^2_2 = 316{,}388$ avec un nombre de degrés de liberté $ddl'_2 = 300$ (degré de signification $p'_2 = 0{,}247$).

VARIATION DE L'EFFET DES VARIABLES EXPLICATIVES SELON LES CATEGORIES DES VARIABLES ALCOOL.

Dans le but de vérifier la validité du modèle à catégories adjacentes, un modèle plus général appelé

modèle « nominal » a été considéré où les odds-ratios peuvent varier selon la catégorie de la variable expliquée.

Plus précisément ce modèle, pour la consommation d'alcool par exemple, suppose, pour un sujet présentant le vecteur de covariables X :

$$\log\frac{\Pr(pb)}{\Pr(nb)} = \alpha_1 + \beta_1 X,\ \log\frac{\Pr(mb)}{\Pr(nb)} = \alpha_2 + \beta_2 X \quad \text{et} \quad \log\frac{\Pr(gb)}{\Pr(nb)} = \alpha_3 + \beta_3 X.$$

$$\log\frac{\Pr(pb)}{\Pr(nb)} = \alpha_4 + \beta_4 X,\ \log\frac{\Pr(mb)}{\Pr(pb)} = \alpha_5 + \beta_4 X \quad \text{et} \quad \log\frac{\Pr(gb)}{\Pr(mb)} = \alpha_6 + \beta_4 X.$$

Il s'ensuit que, si ce dernier modèle est supposé vrai, les paramètres du modèle nominal vérifient :

$\beta_2 - \beta_1 = \beta_3 - \beta_2 = \beta_4$.

Les égalités ci-dessus définissent l'hypothèse nulle H_0 à tester dans le but de rejeter éventuellement le modèle à catégories adjacentes en faveur du modèle nominal.

Pour tester H_0, la statistique du log du rapport des vraisemblances peut être utilisé ; elle suit sous H_0 une loi de chi-deux à quatre degrés de liberté.

En effet, l'hypothèse H_0 exprime deux contraintes sur les six paramètres du modèle nominal, d'où le nombre de degrés de liberté égal à quatre.

TABLEAU III. Odds-ratio entre la consommation d'alcool et les variables explicatives (modèle du type nominal) .

Groupe nb

Variable	Odds-ratio	Intervalle de confiance à 95%	Degré de signification
Sexe	**5**	**(1,90 – 13)**	**$P < 10^{-4}$**
Matripa	**0,086**	**(0,01 – 0,71)**	**P < 0,02**
Depress	**0,56**	**(0,42 – 0,75)**	**$P < 10^{-4}$**
Social	**0,35**	**(0,25 – 0,49)**	**$P < 10^{-9}$**

Groupe pb

Variable	Odds-ratio	Intervalle de confiance à 95%	Degré de signification
Sexe	**3,10**	**(1,30 – 7,40)**	**$P < 10^{-2}$**
Matripa	**0,15**	**(0,019 – 1,20)**	**P < 0,07**
Depress	**0,62**	**(0,49 – 0,78)**	**$P < 10^{-4}$**
Social	**0,42**	**(0,31 – 0,57)**	**$P < 10^{-8}$**

Groupe mb

Variable	Odds-ratio	Intervalle de confiance à 95%	Degré de signification
Sexe	**1,20**	**(0,47 – 3,10)**	**$P < 10^{-5}$**
Matripa	**0,13**	**(0,62 – 1,00)**	**P < 0,05**
Depress	**0,80**	**(0,62 – 1,00)**	**P < 0,08**
Social	**0,57**	**(0,42 – 0,79)**	**$P < 10^{-3}$**

Les résultats montrent que les odds-ratios entre les catégories de la variable expliquée « alcool » (nb, pb, mb, gb) et la variable sexe varient très fortement dans une proportion d'environ 1 à 5. On peut noter aussi que la signification associée à la variable « dépress » est plus faible dans la comparaison (mb/gb), ($P < 0,08$), que dans les comparaisons (nb/gb) et (pb/gb) ($P < 10^{-4}$ dans les deux cas).

Les autres significations ne varient pas de façon notable selon les catégories de buveurs.

Le log de vraisemblance de ce modèle est : $L'_3 = -450,135$.

DISCUSSION.

Comparaison des modèles à partir des résultats précédents pour la variable « alcool ».

Les résultats des tableaux I et II montrent que les effets des variables « sexe », « social » et « dépress » sont significatifs et équivalents dans les deux modèles.

La variable « matripa » a un effet significatif moins marquant mais identique dans les deux modèles.

Par contre, on remarque que l'effet de la « dépressivité » et dans une moindre mesure du sexe varie très fortement dans les catégories de buveurs (voir tableau III).

La comparaison des modèles à catégories adjacentes et nominal à partir du log du rapport de vraisemblance montre que la valeur est significative pour une distribution de chi-deux à quatre degrés de liberté ($P<0,05$). Le modèle nominal peut donc être choisi plutôt que le modèle à catégories adjacentes.

Enfin, la comparaison des résultats obtenus avec les modèles à catégories adjacentes et cumulatif, comme dans le cas de la consommation d'alcool, ne permet pas de préférer un modèle plutôt qu'un autre.

CONCLUSION

Le but de ce travail était de comparer les différents modèles logistiques à logit cumulatif, logit à rapport continu et logit à catégories adjacentes.

Nous avons interprété les différents modèles en termes de distribution de la variable expliquée conditionnellement aux variables explicatives.

Le modèle cumulatif est préférable dans le cas où ces distributions sont translatées l'une de l'autre entre les différents niveaux des variables explicatives.

Le rapport à rapport continu convient mieux si ces distributions se déduisent l'une de l'autre par changement d'échelle.

Pour le modèle à catégories adjacentes, nous n'avons pas pu mettre en évidence une telle distribution qui serait associée au modèle.

Dans notre application, il n'est pas apparu de différences notables entre les modèles à catégories adjacentes et cumulatif et il semblait difficile de comparer les distributions de la variable expliquée « consommation d'alcool » dans les différentes catégories de variables explicatives. En effet, les variables « consommation d'alcool » ne comprenaient que quatre modalités. D'autre part à notre connaissance, il n'existait pas dans la littérature, d'indications à ce sujet.

En conclusion, le choix de l'un des deux modèles peut être effectué au vu des distributions de la variable expliqué quand celle-ci comprend un nombre suffisant de modalités et quand les variables explicatives sont en nombre modéré.

Le choix peut aussi, éventuellement être effectué en fonction des connaissances à priori sur la distribution de la variable expliquée.

Ces deux modèles ont été comparés car ils existaient dans le logiciel BMDP dont l'utilisation est très répandue.

REFERENCES

A. Agresti, « *Analysis of Ordinal Data* », John Wiley and Sons Inc., 1984.

(1) A. Agresti, « Categorial Data Analysis », John Wiley and Sons Inc., 1991.

(2) M. Choquet, S. Ledoux et H. Menke, « *Approche longitudinale des consommations de drogues et des troubles somatiques et psychosomatiques* », INSERM, La Documentation Française, Paris, 1988.

(3) Logiciel BMDP, *Statistical Software*, Volume 2, University of California, Los Angeles Press, 1988.

(4) A. Agresti, « *Categorial Data Analysis* », John Wiley and Sons Inc., 1990.

(5) L.A. Goodman, « The Anlysis of Dependance In Cross-classification having Ordered Categories using log-linear models for frequencies and log-linear models for odds », *Biometrics* , 39, 1983, pp.149-160.

(6) J. Snell, « *A Scaling Procedure for Ordered Categorial Data* », *Biometrics* , 20, 1964, pp.592-607.

(7) O.D.. Williams and J.E. Grizzle, « Analysis of Contingency Tables having Ordered Responses Categories », *J. Amer. Statist. Assoc*., 67, 1972, pp.55-63.

(8) M.C. Cullagh, « Regression Models for Ordinal Data (With discussion) », *J. Roy. Statist. Soc*., B.42, 1980, pp.109-142.

(9) R.D. Bock, « Multivariate Statistical Methods in Behavioral Research », New York, Mc Graw Hill, 1975.

(10) W.A. Thompson, « On The Treatment of Grouped Observations in Life Studies », *Biometrics* , 33, 1977, pp.463-470.

(11) D.R. Cox, « Regression Models and Life Tables (With discussion) », *J. Roy. Statist. Soc*., B.34, 1972, pp.187-220.

(12) S.E. Fienberg and W.M. Mason, « Identification and Estimation of Age Period Cohort Models in the Analysis of Discret Archival Data », *Sociological Methodology*, San Francisco, Jossey Bass, 1979, pp. 1-67.

(13) M.C. Cullagh and J. Nelder, « Generalised Linear Models », London Chapman and Hall, 1983.

(14) R.D.Bock and G. Yates « Multiqual log-linear Analysis of Nominal or Ordinal Qualitative by the Method of Maximum Likelihood », Chicago, International Educational Services, 1973.

(15) A. Agresti, « A Survey of Strategies for Modeling Cross-Classification having Ordinal Variables », *J. Amer. Statist. Assoc*., 78, N° 381, 1983.

(16) A. Agresti, « A Survey of Models for Repeated Ordered Categoriel Response Data », *Statistics in Medecine,* Vol.8, 1989, pp.1209-1244.

(17) S.E. Fienberg and W.M. Mason « Identification and Estimation of Age Period Cohort Models in the Analysis of Discret Archival Data », *Sociological Methodology,* San-Francisco, Jossey-Bass Ed., 1989, pp.1-67.

(18) D.G. Clayton, « Some Odds-ratio Statistics for the Analysis of Ordered Categorical Data », *Biometrika*, 61, 1974, pp.525-531.

(19) J. Hundrickx, H.Ganzebook « Occupational Status Attainment in the Netherlands 1920-1990: A Multinomial Logistic Analysis », *European Sociological Reviews, 14,* 1998, pp.387-403.

(20) S. Hoffman and G. Duncan, « A Comparaison of Choice-based Multinomial and Nested Logit models: the Family Structure and Welfare Use Decisions of Divorced or Separeted Women », *Journal of Human Ressources,* 23, (4), 1988, pp.550-562.

(21) D. Mc Faden, « Regression Based Specification Tests for the Multinomial Logit Models », *Journal of Econometrics*, 34, 1987, pp.63-82.