INTERPRETATION DU MODELE LOGISTIQUE A RAPPORT CONTINU EN TERME DE DISTRIBUTION

Reçu le 28/01/2001 – Accepté le 13/11/2002

Résumé

L'objectif de ce travail est consacré principalement à la présentation de deux modèles logistiques à rapports continus pour une variable réponse ordinale à *c* catégories et de leurs interprétations en termes de distribution .

***Mots clés**: Modèle logistique, logit à rapport continu, odds-ratio.*

Abstract

The objective of this study reviews two continuation ratio logit models for ordinal response variable of *c* categories and their interpretation in terms of distribution.

***Key words**: Logistic regression model, continuation ratio-logit, odds-ratio.*

M. CHIKHI-BOUCHOUL
Faculté des Sciences
Département de Mathématiques
Université Mentouri
Constantine(Algérie)

T. MOREAU, M. CHAVANCE
INSERM, Unité 472
Ville-juif Cedex, (France)

C. HUBER, B. BRU
Université René Descartes, Paris V
U.E.R. de Statistiques Appliquées
Paris Cedex 06 (France)

ملخص

يتناول هذا البحث تقديم نموذجي " لوجيستيك" ذات نسب مستقرة لمتغيرة مرتبة ذات" س" أصناف و مناقشتها بصفة توزيعية.

***الكلمات المفتاحية:** نموذج " لوجيستيك" ، متغيرة مرتبة ، إحصاء.*

La régression logistique connaît de nos jours un usage croissant en épidémiologie [1,2], objectivée par son intégration récente dans certains logiciels statistiques tel que les logiciels SAS et BMDP [3], cette modélisation statistique permet d'estimer l'effet des facteurs de risques sur la maladie et évaluer l'existence d'une relation entre maladie et exposition [4].

Le but de cet article est de rappeler les deux modèles logistiques à rapport continu [5] puis de commenter leurs interprétations, en terme de distributions, 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.

MODELE LOGISTIQUE GENERAL

Le modèle logistique général exprime le logit correspondant à une variable réponse binaire *Y* qualitative en fonction des variables explicatives *X*, qui peuvent être qualitatives ou quantitatives [1,6], il est défini par :

$$\log\frac{\pi(X)}{1-\pi(X)}=\alpha+\beta_1' x_1+...+\beta_k' x_k$$

où $X(x_1...x_k)$ est un vecteur de *k* covariables et β' est un vecteur de *k* paramètres inconnus.

Le modèle envisagé généralise le précédent au cas où la variable réponse comprend *c* catégories ordonnées [5,7].

L'objectif de cette étude consiste particulièrement à présenter l'interprétation de deux modèles logistiques en termes de distribution dans le cas où le modèle est discret [8,9].

Ces deux modèles sont les modèles à rapports continus. Ils sont formés selon l'ordre des c catégories de la variable réponse.

FORMULATION DES LOGITS POUR UNE REPONSE ORDINALE A c CATEGORIES

Les modèles logistiques à étudier expriment le logit correspondant à une variable réponse à c catégories ordonnées ($c > 2$) en fonction de variables explicatives [5,10].

Ces logits utilisent les probabilités π_j que la variable réponse Y appartienne à la $j^{ème}$ catégorie avec $j = 1,\ldots,c$.

Le premier modèle est défini par :

$$L_j = \log \frac{\pi_{j+1}}{\pi_1 + \ldots + \pi_j} \text{ avec } j = 1,\ldots,c-1$$

L_j est ici le logit de $\dfrac{\pi_{j+1}}{\pi_1 + \ldots + \pi_{j+1}}$

Le deuxième modèle est donné par :

$$L'_j = \log \frac{\pi_j}{\pi_{j+1} + \ldots + \pi_c} \text{ avec } j = 1,\ldots,c-1$$

L'_j apparaît comme le logit de $\dfrac{\pi_j}{\pi_j + \ldots + \pi_c}$

Soit X un vecteur de k variables explicatives, le modèle général conditionnellement à X est donné par :

$$L_j(X) = \alpha_j + \beta' X$$

où β' un vecteur de paramètres inconnus $(\beta_1 \ldots \beta_k)$ à estimer ainsi que les α_j, $j = 1,\ldots,c-1$.

On utilise les modèles de régression logistiques à rapport continu dans le cas de l'analyse des données ordonnées discrètes et dans les données de survies [11,12].

ESTIMATION DES PARAMETRES DU MODELE LOGISTIQUE

Les paramètres du modèle logistique général peuvent s'obtenir à partir de (c-1) régressions logistiques en comparant les sujets d'une catégorie à ceux des catégories supérieures.

On montre que les résultats sont asymptotiquement équivalents à ceux d'une analyse unique [13,14].

Cette approche se fait en utilisant un logiciel de régression logistique pour l'analyse des données binaires .

L'estimation des paramètres dans le modèle logistique général [12] :

$$L_j(X) = \alpha_j + \beta' X$$

se fait par la méthode du maximum de vraisemblance [2,15].

Le principe de la méthode du maximum de vraisemblance est de choisir pour estimateurs de α et β les valeurs qui rendent V la vraisemblance de l'ensemble des observations maximum [2].

TEST DE SIGNIFICATION DE PARAMETRES

Le test de signification des paramètres utilisé dans l'hypothèse portant sur la nullité d'un ou plusieurs paramètres du modèle logistique général est le test du log du rapport des vraisemblances, dans le cas des variables explicatives catégorielles [13], le test est analogue au test du χ^2. Il s'interprète de la même manière en termes d'indépendance conditionnelle entre les variables explicatives et la variable réponse [16,17].

Le logiciel statistique utilisé pour vérifier la validité de cette hypothèse est le BMDP [3].

INTERPRETATION DES MODELES LOGISTIQUES AVEC VARIABLE REPONSE ORDINALE

Dans ce paragraphe, les hypothèses du modèle logistique sont explicitées en termes de distribution de la variable réponse Y conditionnelle aux variables explicatives.

Deux modèles particuliers seront considérés, correspondant au logit à rapport continu [5,6].

Définitions et notations

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 $a_0 < a_1 <,\ldots,< a_c$ les limites (observées ou non) des c intervalles contigus adjacents où Y prend ses valeurs, la probabilité que Y appartienne à la classe j sera :

$$\pi_j = \Pr(a_{j-1} < Y < a_j)$$

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

$$F(a_j) = \Pr(Y \le a_j) = \pi_1 + \ldots + \pi_j$$

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.

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

$$Fx(a_j) = \Pr(Y \le a_j / X = x)$$

et $Sx(a_j) = 1 - Fx(a_j)$ est la fonction de survie correspondante.

Les modèles logits à rapports continus

Dans ce paragraphe, la démarche consistera à trouver la limite continue du modèle discret défini pour les variables catégorielles.

Deux modèles différents à logits à rapport continu, peuvent être envisagés. En notant :

$$p_j = \Pr\left[(a_j < Y \le a_{j+1}) / (Y > a_j)\right]$$

et

$$p'_j = \Pr\left[(a_j < Y \le a_{j+1}) / (Y < a_{j+1})\right]$$

ils expriment le logit de p_j ou de p'_j conditionnels à $X = x$ sous la forme :

$$L_j(x) = \alpha_j + \beta x$$

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 [11,18].

Dans ce cas :

$$p_j = \Pr(a_j < Y \le a_{j+1}) / S(a_j)$$

tend vers $f(a_j)/S(a_j)dy = \lambda(a_j)dy$

où $\lambda(a_j)$ est par définition la fonction de "risque instantané" de Y en a_j.

De plus $(1-p_j)$ tend vers 1.

De la même façon :

$$p'_j = \Pr(a_j < Y \le a_{j+1}) / F(a_{j+1})$$

tend vers $f(a_j)/F(a_j)\ dy$ où $f(a_j)$ est la densité de probabilité de Y en a_j.

De plus, $(1-p'_j)$ tendent 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 limites s'écrivent respectivement :

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

En écrivant le premier modèle pour x_1 et pour x_2, on obtient :

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

Soit après intégration :

$$Sx_2(a_j) = \left\{Sx_1(a_j)\right\}^{\exp\beta(x_2 - x_1)}$$

Le deuxième modèle conduirait de même à :

$$Fx_2(a_j) = \left\{Fx_1(a_j)\right\}^{\exp\beta(x_2 - x_1)}$$

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

Le premier modèle est connu sous le nom de "modèle de risques instantanés proportionnels" ou modèle de Cox [4] et il est couramment utilisé dans les études de survie. Un cas particulier de ce modèle est celui où :

$$S_x(y) = \exp\left\{-\lambda y \exp(\beta x)\right\}$$

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 :

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

En effet, en écrivant :

$$Fx_1(a_j) = \left(\frac{a_j}{k}\right)^{t\exp\beta x_1}, \quad Fx_2(a_j) = \left(\frac{a_j}{k}\right)^{t\exp\beta x_2}$$

le modèle est vérifié.

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 les modèles à rapport continu si, entre les différentes catégories de la variable X la variable Y subit un changement d'échelle.

Si la réponse est considérée comme réellement qualitative, ce modèle a une interprétation bien adaptée aux processus unidirectionnels . La probabilité d'être au stade j d'une maladie sachant que le stade j-1 à été dépassé.

Thomson à utilisé ces logits dans les modèles de l'analyse discrète (survival – time data) [18].
Fienberg et Masson ont utilisés le logit à rapport continu dans les modèles (age -période - cohorte) [11].

CONCLUSION

Le but de ce travail est d'interpréter les différents modèles en terme de distribution de la variable expliquée conditionnellement aux variables explicatives.

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

REFERENCES

[1]- Bouyer J., "La Régression Logistique en Epidémiologie", Partie I, *Rev. d'Epidémiologie et de Santé Publique*, 39, (1991), pp. 79-87.

[2]- Bouyer J., "La Régression Logistique en Epidémiologie", Partie II, *Rev. d'Epidémiologie et de Santé Publique*, 39, (1991), pp. 183-196.

[3]- BMDP Statistical Sofware, Vol. 2, University Of California Press, Los Angeles, (1988).

[4]- Cox D.R. and Snell E.J., "Analysis of Binary data", London Chapman and Hall, (1989).

[5]- Agresti A., "Categorical Data Analysis", John Wiley and Sons Inc., (1991).

[6]- Anderson J.A., "Regression and Ordered Categorical Variables", *Journal of the Royal Statistical Society*, Series B46, (1984), pp. 1-30.

[7]- Goodman L.A., "The analysis of Dependence in Cross-Classification having Ordered Categories using log-linear Models for Frequencies and log-linear Models for Odds", *Biometrics*, 39, (1983), pp. 149-160.

[8]- Williams O.D., and Grizzle J.E., "Analysis of Contingency Tables having Ordered Response Categories", *J.Amer .Statist. Assoc.*, 67, (1972), pp. 55-63.

[9]- Cullagh M.C., "Regression Models for Ordinal Data", (With discussion), *J. Roy. Statist. Soc*., B42, (1980), pp.109-142.

[10]- Bock R.D., "Multivariate Statistical Methods in Behavioral Research", New York, Mc Graw-Hill. (1975), p. 29.

[11]- Fienberg S.E. and Mason W.M., "Identification and Estimation of Age Period Cohort Models in the Analysis of Discrete Archival Data", *Sociological Methodology*, San-Francisco, Jossey-Bass, (1979), pp.1-67.

[12]- Mc Cullagh P. and Nelder J., "Generalized Linear Models", Chapman and Hall, London, (1983).

[13]- Back R.D. and Yates G., "Multiqual, log-linear Analysis of Nominal or Ordinal Qualitative by the Method of Maximum Likelihood", Chicago, International Educational Services, (1973).

[14]- Hanfelt J.J. and Liang K-Y, "Approximate Likelihood Ratios for General Estimating Functions", *Biometrika*, 82, (1995), pp.461-477.

[15]- Moreau T., O'Quigley J. and Mesbah M., "A Global Coodness-of fit Statistic for the Proportional Hazards Model", *The Journal of the Royal Statistical Society*, Vol. 34, N°3, (1985), pp. 212-218.

[16]- Bouchoul-Chikhi M., Moreau T., Chavance M. et Bru B., "Modèle logistique cumulatif pour une Variable Réponse Ordinale à *c* Catégories", *Rev. Sc. et Tech*., *Université Mentouri, Constantine*, N°11, (1999), pp. 17-19.

[17]- Hendrix J., Ganzeboom H.B.G.,"Occupational Status Attainment in the Netherlands,1920-1990, "A Multinomial Logistique Analysis", *European Sociological Review*,14, (1998), pp.387-403.

[18]- Crouchley R., "Arandom Effects Model For Ordered Categorical Data", *J. Am. Stat. Assoc*., (1995), pp. 489-498.

□