Essais de modélisation de l'épilepsie en Tunisie: Théorie et application basées sur des modèles de régression logistique

Abdelwaheb Daouthi, Mohamed Dogui, Abdeljelil Farhat

Essais de modélisation de l'épilepsie en Tunisie : Théorie et application basées sur des modèles de régression logistique

Abdelwaheb Daouthi* & Mohamed Dogui** & Abdeljelil Farhat*

** Unité de recherche EAS-Mahdia*
Faculté des sciences économiques et de gestion de Mahdia,
Université de Monastir, Tunisie.
*** Laboratoire de Physiologie, Faculté de médecine de Monastir,*
Université de Monastir, Tunisie.

Résumé

Le but de ce travail est de faire une modélisation de l'épilepsie en Tunisie. Nous avons essayé à travers cette modélisation de montrer la pertinence des modèles de régression logistique comme outils d'analyse. Pour atteindre cet objectif, nous avons commencé par la présentation des modèles de régression logistique. Ensuite, nous avons étudié la méthode d'estimation du maximum de vraisemblance. Enfin et après avoir construit une base de données, nous avons estimé plusieurs modèles de régression logistique appliqués à un groupe de variables quantitatives et qualitatives relatives à 5000 patients épileptiques. Les principaux de ces modèles sont le logit et le probit multinomiaux. L'estimation de ces modèles et l'analyse des résultats obtenus nous ont permis de montrer dans quelle mesure les paramètres sélectionnés sont significatifs.

Mots clés: régression logistique; modèle logit multinomial; modèle probit multinomial; estimation du maximum de vraisemblance; épilepsie.

Abstract

The aim of this work is to make a model of epilepsy in Tunisia. We have tried through this modelling to show the relevance of logistic regression models as analytical tools. To achieve this goal, we have started with the presentation of logistic regression models. Next, we have studied the estimation method of maximum likelihood. Finally, and after building a database, we estimated several logistic regression models applied to a group of quantitative and qualitative variables for 5000 patients with epilepsy. The main models are: the multinomial logit and multinomial probit. The study of these models has enabled us to show to what extent the selected parameters are significant.

Key words: Logistic Regression; Logit multinomial model; Probit multinomial model; Maximum Likelihood Estimator; Epilepsy.

Introduction et sommaire

Historiquement, l'étude des modèles décrivant les modalités prises par une ou plusieurs variables qualitatives date de la période des années 1930-1950. Les travaux les plus marquants de cette époque sont ceux de Bliss (1935) et de Berkson (1944). Ces travaux traitent les modèles dichotomiques simples (modèles logit et probit). Les premières applications ont été alors ménées dans le domaine de la biologie. Ainsi, ce n'est que récemment, que les modèles multinomiaux ont été utilisés. Nous pouvons citer, à titre d'exemple, les travaux de MacFadden et al. (1978) qui ont développé le modèle logit multinomial. Mais, ce modèle présente une principale limite connue sous le nom de l'hypothèse d'indépendance des alternatives non pertinentes. En effet, un autre modèle plus flexible et acceptant n'importe quelle structure d'erreurs a été développé par Daganzo (1979). C'est le modèle probit multinomial.

En tant que procédure non paramétrique, la régression logistique présente l'avantage de ne pas exiger de contraintes quant à la normalité des distributions des variables. La régression logistique est moins une méthode d'inférence statistique qu'une méthode de classification. En effet, l'équation étudiée traduit la probabilité d'appartenance d'un sujet à une catégorie ou un groupe. Tous les modèles logistiques permettent d'analyser des situations réelles. Ils peuvent nous donner des résultats très précieux. Ces modèles sont utilisés surtout pour expliquer certaines maladies qui sont jugées très compliquées (cancer, maladies cardiovasculaires, épilepsie, ...). En Tunisie, si les modèles de régression logistique ont été utilisés dans les études économiques, nous ne les avons pas rencontré dans des recherches médicales ou biomédicales. D'où l'objectif principal de ce travail est d'essayer, à travers une modélisation logistique multinomiale de l'épilepsie en Tunisie, de montrer la pertinence de ce genre d'outils d'analyse. L'épilepsie est définie comme étant une maladie du système nerveux résultant d'un dérèglement passagé de certaines fonctions cérébrales. Notre choix de l'épilepsie est fondé sur les caractéristiques particulières suivantes de cette maladie: cette maladie peut toucher toutes les catégories d'âge; elle ne nécessite pas un traitement particulier et le succès d'un traitement dépend de plusieurs facteurs. Donc, il est nécessaire d'étudier tous ces facteurs qui sont considérés comme déterminant du type de traitement.

Pour répondre à cet objectif, nous avons consacré une première partie à l'étude de différents modèles de la régression logistique binaire à savoir les modèles logit, probit et log-log complémentaire ainsi que les modèles multinomiaux les plus utilisés en pratique qui sont le modèle logit multinomial et le modèle probit multinomial. Les méthodes d'estimation des paramètres des principaux modèles seront abordées aussi dans le cadre de cette recherche, principalement, la méthode du maximum de vraisemblance qui est considérée la méthode d'estimation la plus pertinente de ce type de modèles. En effet,

la partie théorique sera divisée en deux importantes sous-sections. La première porte sur la présentation des principaux modèles de régression logistique. La deuxième est consacrée à l'étude de la méthode d'estimation du maximum de vraisemblance. De plus, nous avons décrit en détail la méthode utilisée pour l'estimation des paramètres des modèles de régression logistique pour la validation ainsi que pour la réalisation de certaines inférences statistiques.

Pour la partie empirique de cette étude, nous avons construit une base de données formée des informations relatives à un échantillon composé de 5000 patients épileptiques de l'hopital de Sahloul (Sousse, Tunisie). Puisque nous sommes les premiers qui ont utilisé ces données, nous avons jugé qu'il est nécessaire d'élaborer des statistiques descriptives sur les différentes variables utilisées dans notre étude. A partir de ces statistiques descriptives, nous avons pu dégagé quelques remarques. D'abord, la maladie d'épilepsie est présente surtout chez les nourrissons et les enfants. En second lieu, nous pouvons dire que cette maladie n'est pas héréditaire du fait que la plupart des épileptiques ne présentent pas d'antécédents. Elle peut toucher toutes les catégories d'âge sans exception. Cette partie de calculs et d'analyses a fait l'objet de la deuxième section de ce travail.

La section 3 contient les résultats expérimentaux d'estimation d'un modèle logit multinomial jugé le meilleur après un nombre élevé d'expériences. Dans cette section, nous commençons par une étude des corrélations entre les variables considérées qui nous a permis la détermination d'un modèle logit multinomial adéquat capable de fournir une meilleure explication de la variable endogène de notre étude à partir de différentes variables explicatives. Cette variable, bien connue en médécine par son nom de Electroencéphalogramme (EEG), est un examen fonctionnel explorant l'activité électrique produite spontanément par les cellules nerveuses. L'étude de corrélations entre la variable expliquée et le grand nombre des variables jugées explicatives nous a permis de conclure que seulement douze des variables indépendantes sont corrélées avec la variable dépendante. Parmi ces variables, nous avons sélectionné seulement sept variables pour être présentées lors de l'estimation du modèle logit multinomial. Puis, nous présentons les résultats relatifs à l'estimation des paramètres. Cette estimation a été suivie par des analyses et des interprétations des rôles des différentes variables du modèle.

La principale caractéristique du modèle logit multinomial est qu'il permet d'estimer $(k-1)$ modèles avec k le nombre de modalités en considérant la modalité manquante comme modalité de référence. Dans notre étude, la modalité de référence est celle relative à la modalité "autres types d'EEG". Après l'estimation et l'analyse des résultats, nous avons pu tirer les principales conclusions suivantes:
- la catégorie d'âge est la variable la plus importante puisqu'elle est présente dans tous les modèles relatifs aux différentes modalités du modèle logit multinomial et
- les variables: "antécédents personnels (nourrisson, enfant)", "délais de la dernière crise"

et "le type de l'EEG" sont aussi d'une grande importance du fait qu'elles sont présentes dans la plupart des modèles. Donc, nous pouvons dire que ces trois variables sont supposées pertinentes quant à l'explication de la variable réponse. Nous nous intéressons ensuite à la variable "type des crises" qui est jugée par les spécialistes de neurologie comme la variable la plus intéressante. Nous remarquons que cette variable n'est incluse dans aucun modèle donc elle n'est pas significative.

Dans le cadre de notre étude, des comparaisons faites entre les modèles logit et probit ont montré que dans la plupart des cas, ces deux modèles dégagent des résultas très proches. Nous avons pu dégager aussi les principaux avantages ainsi que les inconvénients de la régression logistique. Concernant les avantages, nous avons d'abord remarqué que les odds ratio sont faciles à interpréter. De plus la régression logistique permet de traiter des variables explicatives: qualitatives et quantitatives (discrètes ou continues). Elle permet aussi de détecter certains phénomènes non linéaires. D'autre part, il n'y a pas d'hypothèse de normalité ni d'homoscédasticité.

Nous avons souligné aussi quelques inconvénients liés à l'utilisation de ces types de modèles. Premièrement, les modèles de la régression logistique ne s'appliquent qu'aux données sans valeurs manquantes. En second lieu, ils sont sensibles aux observations aberrantes. Enfin, ils supposent la non-colinéarité des variables explicatives.

Finalement, nous pouvons dire qu'il est souhaitable de mener une étude longitudinale afin d'enrichir notre travail.

Bibliographie

[1] Berkson, J. (1944) Application of the logistic function to bioassay. *Journal of The American Statistical Association*, 39: 357-365.
[2] Bliss, C.I. (1935) The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22: 134-167.
[3] Daganzo (1979) Multinomial Probit: theory and its application forecasting. *Academic Press, New York.*
[4] Greene, W.H. (2003) Econometric Analysis. *5e éd., Upper Saddle River, N.J.,Prentice Hall.*
[5] Maddala, G.S. (1983) Limited dependent and qualitative variables in econometrics. *Cambridge university Press, New York.*
[6] Macfadden, D. and Tye W. and Train K., (1978) An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model *Transportation Research Record: Forecasting passenger andfreight travel*, No. 637, 39-46.
[7] Muller, Ch.H. and Neykov, N. (2003) Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, 116, 503-519

[8] Prentice, R.L. (1976) A generalization of the probit and logit methods for dose response curves. *Biometriks*, 32: 761-768.
[9] Reggiani, A. and al. (1989) A new approach to modal split analysis: some empirical results. *Transportation Research*, 23B: 75-82.
[10] Simeckova, M. (2005) Maximum Weighted Likelihood Estimator in Logistic Regression. *WDS'05 Proceedings of Contributed Papers*, Part 1: 144-148.
[11] Vandev, D et Neykov, N. (1998) About Regression Estimators with High Breakdown Point. *Statistics*, 32: 111-129.
[12] Vani K. B. (2002) Logit and probit: ordered and multinomial models. *Quantitative Application in the Social Sciences*, 07-138: 97 pages.
[13] Wald, A. (1941) Asymptotically Most Powerful Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 12: 1-19.
[14] Zhen, C. et Lynn, K. (2001) A Note on the Estimation of the Multinomial Logit with Random Effects. *The American Statistician*, 55 [2]: 89-95.