COMPARISON OF SMOOTHING METHODS IN LOGISTIC REGRESSION MODEL WITH APPLICATION TO BIOCHEMICAL ANALYSIS

Souad Bechrouri [1] & Mohamed Choukri [2] & Abdelilah Monir [3] & Hamid Mraoui [1] & Ennouamane Saalaoui [2]

[1] *Laboratoire de Recherche en Informatiques, Faculté des sciences, Université Mohamed Premier, Oujda, Maroc - souad.bechrouri@gmail.com, hamid_mraoui@yahoo.fr*

[2] *Laboratoire de biochimie,Faculté des sciences, Université Mohamed Premier, Oujda, Maroc - choukrimohamed@hotmail.com, saalaoui_ennouamane@yahoo.fr*

[3] *Laboratoire de Modélisation Stochastique et Déterministe,Faculté des sciences, Université Mohamed Premier, Oujda, Maroc - Abdelilah.monir@gmail.com*

Résumé. La régression logistique est l'un des modèles les plus utilisés en études cliniques et épidémiologiques. Elle permet d'estimer la relation entre une variable dépendante binaire et une ou plusieurs variables explicatives catégorielles ou continues. Concernant ce dernier type, les spcialistes ont souvent recours il est fréquent de le transformer en variable catégorielle; cependant cela induit une perte d'information. Par ailleurs, la relation entre les variables explicatives et la variable sortie est une relation linaire. Dans cette étude, l'effet non-linéaire est estimé par des fonctions splines afin d'ajouter plus de flexibilité. l'objectif est de modéliser la probabilité d'avoir une hyperglycémie, chez les patients hospitalisés au Centre Hospitalier Universitaire Mohamed VI d'Oujda, par quelques paramètres biochimiques et l'âge. Nous avons comparé des techniques utilisant des fonctions splines pour ajuster l'effet des variables continues dans un modèle de régression logistique qui sont : les splines pénalisées, les splines naturelles et les splines cubiques restreintes. Pour évaluer ces méthodes, nous utilisons la déviance résiduelle. Nous observons que les splines pénalisées ajuste au mieux le modèle avec la plus petite déviance résiduelle. La sélection des variables significatives dans le modèle de régression est effectuée par la méthode bidirectionnelle. Les analyses statistiques sont réalisées avec la version 3.2.3 du logiciel R. Le niveau de signification est considéré à 5%.

Mots-clés. spline, hyperglycémie, régression logistique, ...

Abstract. Logistic regression is a popular used models in epidemiologic and clinical studies. It was been used to estimate the relation between dependent binary variable and independent categorical and/or continuous variables. In this latter, the transformation to categorical variable has been frequently used; however, it could cause a loss of information. In this study, the nonlinear effect is fitted by spline functions to allow the flexibility. Further, the interest is to model the probability to have hyperglycaemia event in hospitalized patients in the University Hospital Center (CHU) Mohamed 6 Oujda by some biochemical parameters and age. In this paper, we compared techniques using spline functions for

adjusting the effect of continuous variables in logistic regression model namely penalized spline, natural spline and restricted cubic spline. To evaluate these methods, we request the residual deviance (log likelihood). The results showed that penalized splines are the best model with the lowest residual deviance. The selection of significate variables in the regression model was defined by the stepwise (both) selection procedure. Statistical analyses were performed with the R 3.2.3 Software and the significant level is considered at 5 %.

Keywords. spline, hyperglycaemia, logistic regression ...

1 Introduction

The logistic regression is commonly used model in epidemiological and clinical studies for evaluating the effect of independent continuous or categorical variables (X) on dependent binary variable (Y). This model predicts the probability to have an event. It is largely used due to its simplicity. Otherwise, this model is based on linearity assumption between the probability to have an event and predictors. When the explanatory variable is continuous, it is frequently transformed into categorical variable especially in epidemiological studies (Turner, Dobson, & Pocock, 2010). The advantage of this transformation is the simple interpretation of results and graphical representation. However this transformation induces a loss of power, it is ineffective fitting, and a poor modelling of the relationship between X and Y. So, researchers have proposed splines functions to avoid these problems. Splines functions can be used as an alternative to converting continuous to categorical variables (Harrell, 2006).
Desquilbet & Mariotti, (2010) have used restricted cubic spline to check the linearity assumption. In this study we search the non-linear effect of different biochemical parameters and age on the hyperglycaemia status.
This paper is organized as follows: in Section 2, we describe the data set and the statistical method performed. In Section 3, we present the comparative evaluation via various experiments and discuss the findings. We conclude this paper in Section 4.

2 Data and Statistical Tools

2.1 Dataset

Dataset has been included hospitalized patients of the CHU Mohamed six, Oujda. The interest is to model the hyperglycaemia status by age, triglyceride level, albumin, chlorine, HDL cholesterol and urea. The age of patients is ranged from 20 to 90 years. The hyperglycaemia status has been defined on the Fasting Blood Sugar (FBS) registered in the first admission when the FBS was upper than 1.26 g/L. All the rest biochemical analysis has been completed in the biochemical laboratory of CHU Oujda. The original

data set contains a lot of missing data, and at the same time does not have all the interesting variables to model hyperglycaemia. In this work, we are based on complete case study, so the rows are removed if they contain any missing value.

2.2 Methods and tools

To avoid the linearity assumption in the logistic model, the use of spline is requested. In this section, we present three non-parametric smoothing methods that are investigated: penalized spline, natural spline and restricted cubic spline. Statistical analysis has been performed by R 3.2.3 tool and error level is considered at 5%.
The parametric logistic regression function is defined as:

$$P(Y|X_1,...,X_n) = \frac{1}{(1+\exp\left(-\left(\alpha+\sum_{i=1}^{n}\beta_i X_i\right)\right))} \tag{1}$$

In this model, continuous variables are modelled as linear regression. To model the non-linearity relationship between output and input variables, spline functions are requested. They are the piecewise polynomials functions which are connected by knots. They have been used primarily in physical sciences. The simple form spline was the linear spline function, a piecewise linear functions. Splines of order d is a linear combination of basis functions or piecewise polynomial functions of order d. They were (d -1)-times continuously differentiable at the knots. Basis function can be truncated power basis or B-splines. Truncated power basis are defines as below:

$$x_{min} \leq \kappa_1 < ... < \kappa_k \leq x_{max}$$

$$f(x) = \alpha + \sum_{i=1}^{d}\alpha_i x^i + \sum_{j=1}^{k}\gamma_j (x-\kappa_j)_+^d \tag{2}$$

Where

$$(x-\kappa_j)_+ = \begin{cases} x-\kappa_j & if x > \kappa_j \\ 0 & if x \leq j \end{cases}$$

The B-spline basis (De Boor C. 1978) function are denoted B_j^d. They are defined recursively as

$$B_j^d(x) = \frac{x-\kappa_j}{(\kappa_{j+d-1}-\kappa_j)}B_j^{d-1}(x) - \frac{x-\kappa_{j+d}}{(\kappa_{j+d}-\kappa_{j+1})}B_{j+1}^{(d-1)}(x) \tag{3}$$

Where

$$B_j^1(x) = (\kappa_{j+1}-x)_+^0 - (\kappa_j - x)+^0 = \begin{cases} 1 & for \quad \kappa_j \leq x < \kappa_{j+1} \\ 0 & else \end{cases}$$

The function estimation of *logit* $E(Y|X)$ is $f(x) = \alpha_0 + \sum_{j=-(d-1)}^{m}\gamma_j B_j^d(x)$ where Y_j are estimated for a given sample (y_i, x_i),$i = 1, ..., n$. This expression is given for one covariate. In multivariate regression we use additive spline. The choice of order d and the knot

sequence κ is very important for obtaining a good fit. Three different approaches were used for estimating the non-linear effect for continuous variables.

Penalized spline
For fitting the regression function with penalized spline we search to minimize the penalized residual sum of squares:

$$\min_{\alpha \in R^{m+d}} \sum_{i=1}^{n} \left(y_i - \sum_{j=-(d-1)}^{m} \tilde{\alpha}_j B_j^d(x_i) \right)^2 + \lambda \sum_{-(d-1)+2}^{m} (\triangle \tilde{\alpha}_j)^2 \tag{4}$$

In this quantity Eilers & Marx (1996) have proposed to penalize the high second order differences of the estimated parameters, hence the introduction of P-spline basis.λ is the smoothing coefficient which equilibrate between roughness and smoothness of the regression estimation.

restricted cubic spline Restricted cubic spline is spline cubic with first and second derivatives which are continuous at the inner knots and restrained to be linear in tails (before the first value and after the last). It use truncated power basis.

natural spline
Natural spline is principally a restricted cubic spline which use B-splines basis instead of truncated power basis. It has the following form: $f(x) = \alpha_0 + \sum_{i=1}^{m} \beta_i B_i^d(x)$

3 Application

The aim is to compare different methods of estimation, a logistic regression explaining the probability to have hyperglycaemia by age, albumin, chlorine, HDL cholesterol, urea and triglyceride are used. These variables are selected using the stepwise selection method (both selection method).This method consists on stepwise removal of one predictor from the full model and compares the AIC value by ascendantly ordering them. The process is repeated with the retained model. Terms are subtracted and/or added (both) to allow the comparison of the models. Since the lowest AIC value is still the model. The full model (AIC=491.68) is:

$$\begin{aligned} logit(p) = \alpha_0 + \beta_1 * age + \beta_2 * sex + \beta_3 * service + \beta_4 * albumin + \beta_5 * chlorine + \\ \beta_6 * totalcholesterol + \beta_7 * HDLcholesterol + \beta_8 * potassium + \beta_9 * Sodium \\ + \beta_{10} * urea + \beta_{11} * triglyceride \end{aligned} \tag{5}$$

The both selection is used to obtain this model. Thus we have the restraint model with the lowest AIC. Further, for all analysis the model below is used.

$$logit(p) = \alpha_0 + \beta_1 * age + \beta_2 * albumin + \beta_3 * chlorine + \beta_4 * HDLcholesterol + \beta_5 * urea + \beta_6 * triglyceride \quad (6)$$

The table 1 showed the logistic regression results for the previously cited model. The

Table 1: Results of logistic regression for hyperglycaemia event.

parameters	β	SE	$exp(\beta)$	p_value
Intercept (α_0)	0.2290	2.3108	1.2574 [0.0133-119.0399]	0.9210
Age	0.0269	0.0074	1.0273[1.0127-1.0426]	0.00028
Albumin	0.0748	0.0203	1.0777[1.0378-1.12414]	0.00022
Chlorine	-0.0448	0.0219	0.9560[0.9138-0.9958]	0.04031
HDL cholesterol	-2.1799	0.7979	0.1130[0.0213-0.4900]	0.00629
Urea	0.4959	0.1540	1.6420[1.2234-2.2418]	0.00128
Triglyceride	0.3512	0.146	1.4208[1.0350-1.7523]	0.01639

logistic regression equation is:

$$logit(p) = 0.2290 + 0.0269 * age + 0.0748 * albumin - 0.0448 * chlorine - 2.1799 * HDLcholesterol + 0.4959 * urea + 0.3512 * triglyceride \quad (7)$$

All variables have a significant influence on the probability to have hyperglycaemia in hospitalized patients. The $exp(\beta)$ represents the odds ratio. The odds of having hyperglycaemia are 1.42 times higher among patients with high triglyceride value, for other fixed variables, as compared to patients with normal triglyceride value. The chlorine and HDL cholesterol have a negative influence on hyperglycaemia. These results require a clinical explication. For these variables we have applied a spline transformation. Table 2 presents the penalized deviance of each smoothing method for the model with the previous variables. More the model penalized residual deviance is lower more the model is better. Penalized spline presents the smallest residual deviance value. The others models obtain residual deviance lower than the obtained with simple logistic regression. All variables are significant in the simple model and in models with smoothing methods.

4 Conclusions

Logistic regression using three smoothing techniques for modelling non-linear effect on hyperglycaemia event leads to raise the relationship between age, albumin, urea, chlorine, HDL cholesterol and triglyceride. The result of residual deviance showed that the

Table 2: penalized deviance for each model which explain the probability to have hyperglycaemia.

Model	Penalized Residual deviance
Logistic regression	453.20
Restricted cubic spline	444.07
Penalized spline	429.80
Natural spline	431.80

penalized splines have the lowest value. All basis functions were significant. Otherwise, the two other smoothing methods have a residual deviance lower than the simple model. Previously studies were comparing these smoothing tools for fitting the non-linear effect in Cox model (Roshani & Ghaderi, 2016) but none study for comparing it in the logistic regression. In further work, we will use fractional polynomials and smoothing splines methods in local logistic regression.

Bibliographie

[1] De Boor C. (1978). A Practical Guide to Splines. Springer-Verlag: New York
[2] Desquilbet, L., & Mariotti, F. (2010). Dose-response analyses using restricted cubic spline functions in public health research. Statistics in Medicine, 29(9), 10371057. http://doi.org/10.1002/sim.3841
[3] Eilers, P. H. C., & Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. Statistical Science, 11(2), 89121.
[4] Harrell, F. E. J. (2006). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, 600.
[5] Roshani, D., & Ghaderi, E. (2016). Comparing Smoothing Techniques for Fitting the Nonlinear Effect of Covariate in Cox Models. Acta Informatica Medica, 24(1), 38. http://doi.org/10.5455/aim.2016.24.38-41
[6] Turner, E. L., Dobson, J. E., & Pocock, S. J. (2010). Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. Epidemiologic Perspectives & Innovations, 7, 9. http://doi.org/10.1186/1742-5573-7-9