Logistic Regression and ROC Curves: a Non Conventional Methodology

Haedo, Ana Silvia
University of Buenos Aires, Department of Informatics.
Int. Guiraldes 2160, (1428) Ciudad Universitaria.
Buenos Aires, Argentina.
Haedo@qb.fcen.uba.ar

Natal, Marcela Liliana
University of Mar del Plata, Department of Mathematics
Funes 3350. (7600). Mar del Plata,
Buenos Aires, Argentina
Mnatal@mdp.edu.ar

1. Introduction

The threshold value in laboratory tests used to distinguish between healthy and ill subjects is the most significant decision in the area of clinical diagnosis. ROC (receiver operating characteristic) curves are a useful tool to facilitate taking this decision. A statistical approach to diagnostic tests consists in applying non- linear regression predictive techniques based on logistic distribution. Hui and Miller (1991) used ROC curves to refine the statistical inference of logistic regression models, and applied ROC methodology to response or dependent variables. Up to now, ROC curves are only applied to the dependent or response variable. The co-variables involved in a logistic regression model can be continuos or categorical. When they are continuous, and perhaps due to clinical or epidemiological reasons, they are sometimes categorized for the analysis. In this case, statistical inference is not invariant to the categorization used in the analysis. Our main objective is to propose an unconventional working method to apply ROC curves to the determination of cutoff points on the continuous explicative variables**,** in situations in which the response variable could be analyzed by logistic regression**.**

2. Methodology

ROC curves were used both to define the cutoff point in a continuous variable and to evaluate the ability of the logistic model to distinguish between failure or success based on different probability values like: the area under the ROC curve (Hanley and Mc Neil, 1982), the projected length of the curve (PLC) and the area "swept" by the ROC curve (ASC) (Lee and Hsiao, 1996). The Logistic Regression model was used to describe the relation between one or more explicative, categorical or continuous variables and a dichotomic dependent variable. Information was obtained from a database of HIV patients treated in the Pathology Unit of the Hospital Interzonal of the City of Mar del Plata from 1986 to July 1997.

3. Results

GENDER, RISK1, OMSA1 and AIDS (1: ill, 0: healthy) were the categorical variables, and AGE and CD4A were the continuous variables in the database. Our purpose was to predict AIDS risk fitting a logistic regression model with AIDS as the dependent variable, and GENDER, AGE, CD4A and RISK1 as the independent variables. The model was constructed considering CD4A as continuous variable; then, using ROC curves, a cutoff point was determined in the continuous variable. Its value was 200. Other points (50, 100 and 450), taken from histogram analyses, were

also considered. Different logistic regression models were fitted by the "stepwise" method categirizing variable CD4A for the different cutoff points proposed. In order to choose the model, p value, Hosmer test, the area under the ROC curve (in predicted variable), the percentage of correct classifications, and PLC and ASC indices were taken into account.

TABLE 1: Results of logistic regression models according to discretization variable CD4A

Variable CD4A	FIT p-value	Hosmer	ROC Area	Correct %	$\frac{1}{2}+\frac{\sqrt{2}}{4}$ PLC *	$\frac{1}{2}$ + ASC *
Continuous	0.999	0.206	0.8493	79.1 78.6 ill 79.2 healthy	0.836	0.725
Discrete (50)	0.355	0.827	0.6611	52.7 82.1 ill 44.6 healthy	0.633	0.527
Discrete (100)	0.33	0.482	0.7129	52.7 85.7 ill 43.6 healthy	0.647	0.566
Discrete (200)	0.699	0.334	0.8145	78.3 76.8 ill 78.7 healthy	0.777	0.637
Discrete (450)	0.922	0.807	0.7689	69.4 76.8 ill 67.3 healthy	0.72	0.652

* indices associated to PLC and ASC

The model selected was the one that considered variable CD4A discretized with cutoff point 200.

4. Conclusions

The cutoff point obtained using a ROC curve to categorize variable CD4A agrees with the one used by physicians to determine if a carrier becomes ill (Bartlett, 1998). The methodology proposed was also applied to a study on Chagas' disease in Argentina, with similar results. The different steps of this methodology were: a) the adjustment of the logistic regression model with continuous variable; b) the determination of a cutoff point in the continuous explicative variable by the ROC curve; c) the adjustment of the logistic model using the discretized explicative variable; and d) the comparison between the models obtained applying the statistics of Pearson and Hosmer and Lemeshow, the area under the ROC curve in the predicted variable, the percentage of correct classifications, PLC and ASC indices, and the formulation of the model and the analysis of the estimated coefficients.

We conclude that the ROC curve is an adequate tool to determine a cutoff point in the explicative variable of the Logistic Regression. We propose this research methodology to solve similar specific problems in other research fields.

5. References

Bartlett, J.. 1998. Medical Management of HIV infection. *Published by Johns Hopkins University, Department of infectious diseases. Printed in The United States America by Port City Press*.

Hanley, J.; Mc Neil, B.. 1982. The meaning and use of the area under a ROC Curve. *Radiology*, 143: 29-36.

Hui, S.; Miller, M. 1991.Validaton techniques for logistic regression models. *Statistic in medicine*, 10: 1213-1226.

Lee, W.; Hsiao, C. 1996. Alternative summary indices for the ROC Curve. *Epidemiology* 7:605-611.

Acknowledgements.

This paper was supported by a grant TX097 of the University of Buenos Aires.