ALGORITHM E.M. : MORTGAGE APPLICATIONS

by
SALVADOR JAVIER MOLINA RUIZ
Departamento de Estadística y Econometría
Universidad de Málaga

FRANCISCO CANTALEJO GARCIA
Departamento de Matemáticas
Universidad de Málaga

ABSTRACT

The problem we are analysing arise when we utilise tests "screening" to evaluate the conditions (to diagnose the aptitude) of a number of candidates to receive a mortgage from a financial entity, with the problems of lack of information, more specifically, supposing we have two tests with two possible results respectivally called R^{+} and $R^{-;}$ the results of both tests are conditionally independent and the possibility of obtaining a false positive results for each test is zero. As the objective of this study we propose to utilise a methodological algorithm based on the statistic technique known as E.M., to determine the probability of concession of the credit and to know which test is more reliable.

We present the algorithm E.M. and its relationship to the method of Maximum likelihood and the parametric estimation, and the use of this algorithm to solve the problems which arise when the data is incompleted, censored or manipulated. Consisting of two phases such as complete data relating to data observed or incomplete, following which we define the likelihood functions to which we calculate its average, phase E and then we maximize the above mention value, phase M.

Finally we present a practical case in which we realize all the process of the algorithm E.M., the different expressions obtained being programmed and where apart from some initial values we obtain the different values which give us the probability of concession and so, which is the best test to decide the concession of the mortgages.

INTRODUCTION

The problem we are analyzing arises when we utilize test ”screening” to evaluate the conditions (to diagnose the aptitude) of a number of candidates to receive a mortgage from a financial entity, with the problem of lack of information, for which, and as the objective of the present study we propose the utilization of a methodological algorithm based on the statistics technique known as E.M., that consists of the following:

As we know one of the most important problems in statistics is the point estimation, which is encompassed in the decision making problems and where there are a number of methods to estimate the parameters, one of them being the method of maximum likelihood consisting of the following:

The likelihood function of random variables X_1, X_2, ..., X_n is the join density conjunta) of the n variables g(X_1, X_2, ..., X_n; ?) considered as a function of ?. In particular if $x_1, x_2, ..., x_n$ is a random sample of the random variable X with density function f (X, ?) then the likelihood function will be :

$$L(u; x_1, x_2, ..., x_n) \; ? \quad f(x_{1;} ?)\, f(x_2 ; ?) \ldots f(x_n ; ?) \qquad \text{con } ? \; ? \; W$$

Representing by $y_1, y_2, ..., y_n$ the observed values , we want to know from which density it is more likely that these particular set of values belongs to, so we want to find the value of ?, $\hat{?}$ belonging to the parametric space that will maximize the function of verisimilitude, when that happens we call $\hat{?}$ the maximum likelihood estimator of ?. A lot of the likelihood function fulfil the regularity conditions so the maximum likelihood estimator is the solution of the equation:

$$\frac{? L(? ; X)}{??} \; ? \; 0$$

Because L(?) y log L(?) have their maximum value as the same value of ?, sometimes is more easy and comfortable to calculate the maximum of the logarithm of the likelihood function.

But sometimes the function of verisimilitude is not easy to manipulate, or the data for which we want to calculate the estimation can not be obtained directly , so we have to obtain them through a different set of observed data, “y”, which depends on the value of “x”, we call this problem “incomplete data”, and it can raise problems for the application of the method of maximum likelihood , these inconveniences being the reason why others methods of estimation appear, iterative methods and easy to program, amongst them we point out the algorithm E.M.

In most general cases this algorithm E.M. can be presented as follows:

We consider f (X; ?) a function of density and from this we define :

$$Q\,(?^{'}, ?\,) = E\,(\,\log f\,(\,x\,/\,?\,)\,/\,y\,,\,?\,)$$

We suppose this exists for every pair $(?^{'}, ?)$ and that $f\,(\,x\,/\,?\,) > 0$ in nearly every point of x for every ? of W. We define the E.M. algorithm in the following way :

STAGE E

We calcúlate $Q\,(\,?\,/\,?^{(p)}\,)$

STAGE M

We choose $?^{(p+1)}$ in the way that it will be a value of ? in W, which maximizes $Q(?\,/?^{(p)})$, that means we want to choose $?^{*}$ in a way that maximizes log f(X/?). Because we do not know log f(X/?), instead we maximizes its average given the data “ y “ the value $?^{(p)}$.

APPLICATION

Supposing that we have two tests with two possible results for each of them, R^+ and R^-; and that the results of the two tests are conditionally independent, and that the possibility of a false positive result for every test is almost non-existent (Nil), that means, if an applicant is suitable for obtaining the concession, neither of the tests will have a positive result.

The objective is to estimate the efficiency of each of the tests and the probability of the prevalence of suitability.

To realize the study, the two tests are applied to a random sample of N candidates whose situation is unknown, obtaining the following data:

TABLE 1

Test 2

		R^+	R^-	Total
Test 1	R^+	Y_{11}	Y_{12}	
	R^-	Y_{21}	Y_{22}	
	Total			

This is the data observed which is incomplete data, because the complete data will be X_{11}, X_{12}, X_{21}, X_{221}, X_{222} where it is verified that :

X_{ij} = Y_{ij} except for (i, j) = (2, 2)
X_{221} = nº of unsuitable candidates with (R^+) and (R^-)
X_{222} = nº of suitable candidates

Being ? the probability of the prevalence of suitability, that means, P (E) = ?..

The probabilities P_1 and P_2 measure the efficiency of each test respectively, which means :

$$P_1 = P (R_1^+ / E) \qquad \text{and} \qquad P_2 = P (R_2^+ / E)$$

likelihood function_ of the completed data is therefore the following:

$$L(?,P_1,P_2,X) ? (P_1 P_2)^{X_{11}} (P_1(1?P_2))^{X_{12}} ((1?P_1)P_2)^{X_{21}} ((1?P_1)(1?P_2))^{X_{221}} ?^{N-X_{222}} (1??)^{X_{222}}$$

We calculate sufficient statistics for P_1, P_2, y ?.

Reorganizing the likelihood function we have

$$L(?,P_1,P_2,X) ? P_1^{X_{11}?X_{12}} (1?P_1)^{X_{21}?X_{221}} P_2^{X_{11}?X_{21}} (1?P_2)^{X_{12}?X_{221}} ?^{N-X_{222}} (1??)^{X_{222}}$$

And representing **:**

$X_{1+} = X_{11} + X_{12}$ = nº of unsuitable candidates in the sample with $R_1^?$
$X_{+1} = X_{11} + X_{21}$ = nº of unsuitable candidates in the sample with $R_2^?$
N_E = number of unsuitable candidates in the sample

Then we can express the function of verisimilitude in the following way :

$$L(?\,,P_1,P_2,X)\ ?\ P_1^{X_{1?}}\,(1?P_1)^{N_E\,?\,X_{1?}}\ P_2^{X_{?1}}\,(1?P_2)^{N_E\,?\,X_{?1}}\ ?^{\ N_E}\,(1?\,?\,)^{X_{?1}}$$

Therefore we can state that X_{1+}, X_{+1}, N_E are (sufficient statistics) for P_1, P_2 and ?.

The estimators of maximum verisimilitude are :

$$\hat{P}_1\ ?\ \frac{X_{1?}}{N_E} \qquad\qquad \hat{P}_2\ ?\ \frac{X_{?1}}{N_E} \qquad\qquad \hat{?}\ ?\ \frac{N_E}{N}$$

To calculate these estimators we use the algorithm EM, utilizing the data observed Y_{11}, Y_{12}. Y_{21}, Y_{22}.

Is obvious that this is verified:

$$X_{1+} = Y_{11} + Y_{12}$$

$$X_{+1} = Y_{+1} + Y_{21}$$

The stages of the EM can be easily obtained If $(?^{(k)},\ P_1^{(k)},\ P_2^{(k)})$ represent the estimated parameters after the K-ésima iteration then we will have :

E- STAGE

$$X_{1?}^{(k?1)}\ ?\ E(X_{1?}\,/\ Y,\ ?^{(k)},\ P_1^{(k)},P_2^{(k)})\ ?\ Y_{1?}$$

$$X_{?1}^{(k?1)}\ ?\ E(X_{?1}\,/\ Y,\ ?^{(k)},\ P_1^{(k)},\ P_2^{(k)})\ ?\ Y_{?1}$$

$$N_E^{(k?1)}\ ?\ E(N_E\,/Y,?^{(k)},P_1^{(k)},P_2^{(k)})\ ?\ N\text{-}Y_{22}\,\frac{1??^{(k)}}{(1?P_1^{(k)})(1?P_2^{(k)})?^{(k)}\ ?\,(1??^{(k)})}$$

M- STAGE

$$?^{(k?1)} ? \frac{N_E^{(k?1)}}{N}$$

$$P_1^{(k?1)} ? \frac{Y_{1?}}{N_E^{(k?1)}}$$

$$P_2^{(k?1)} ? \frac{Y_{?1}}{N_E^{(k?1)}}$$

Let us see a practical example:

Supposing that both tests have been elaborated, with the corresponding set of questions and the results presented in table 2 are obtained, where the concession or not of the corresponding mortgages are presented as results, that means in terms of R^+ or R^-.

TABLA 2

Test 2

		+	-	Total
Test 1	+	20	5	25 = Y_{1+}
	-	15	60	
	Total	35 = Y_{+1}		

In view of this data we take as initial values for N$_E$, P_1, P_2 respectively

$N_E^{(0)}$? 40 ?$^{(0)}$? 0.4

$P_1^{(0)}$? $\frac{25}{40}$? 0.625 $P_2^{(0)}$? $\frac{35}{40}$? 0.875

The values obtained in the successive iterations are:

Data observed : 20 5 15 60

Initial values : ?$^{(0)}$? 0.4 $P_1^{(0)}$? 0.625 $P_2^{(0)}$? 0.875

? (1) = 0.4182	P_1(1) = 0.5978	P_2 (1) = 0.8370
? (2) = 0.4270	P_1(2) = 0.5855	P_2 (2) = 0.8197
? (3) = 0.4317	P_1(3) = 0.5792	P_2 (3) = 0.8108
? (4) = 0.4342	P_1(4) = 0.5758	P_2 (4) = 0.8061
? (5) = 0.4356	P_1(5) = 0.5739	P_2 (5) = 0.8034
? (6) = 0.4364	P_1(6) = 0.5728	P_2 (6) = 0.8019
? (7) = 0.4369	P_1(7) = 0.5722	P_2 (7) = 0.8011
?(8) = 0.4372	P_1(8) = 0.5719	P_2 (8) = 0.8006

Final result :

? = 0.4373 P_1 = 0.5717 P_2 = 0.8004

BIBLIOGRAFIA

BLIGHT. B. J. “ ESTIMATION FROM A CENSORED SAMPLE FOR THE EXPONENTIAL FAMILY”. BIOMETRIKA (1972).

BOYLES RUSSELL A. “ ON THE CONVERGENCE OF THE E.M. ALGORITM “(1985).

CLAYTON D, & CUZICK J. “THE E.M. ALGORITMS FOR COX’S REGRESSION” (1994).

DAY N. E. “ ESTIMATING THE COMPONENTS OF A MIXTURE OF NORMAL DISTRIBUTIONS”. (1971).

LAIRD N.M. “EMPIRICAL BAYES METHODES FOR TWO-WAY CONTINGENCY TABLES” (1986).

RUBEN D.E. THAYLER D. “E.M. ALGORITHMS FOR M.L. FACTOR ANALISYS” PSYCOMETRIKA 47. (1982)