Credit Risk Prediction based on Bayesian estimation of logistic regression model with random effects

Mestiri, Sami and Farhat, Abdejelil

Faculty of Management and Economic Sciences of Mahdia

2018

Credit Risk Prediction based on Bayesian estimation of logistic regression model with random effects

Sami Mestiri and Abdeljelil Farhat

Applied Economics and Simulation,
Faculty of Management and Economic Sciences of Mahdia.
Monastir University, Rue Ibn Sina Hiboun, Mahdia Tunisia

Abstract : The aim of this current paper is to predict the credit risk of banks in Tunisia, over the period (2000-2005). For this purpose, two methods for the estimation of logistic regression model with random effects: Penalized Quasi Likelihood (PQL) method and Gibbs Sampler algorithm are applied. By using information on a sample of 528 Tunisian firms and 26 financial ratios,we show that Bayesian approach improves the quality of model predictions in terms of good classification as well as by the ROC curve result.

Key words: Forecasting, Credit risk, Penalized Quasi Likelihood, Gibbs Sampler, Logistic regression with random effects, Curve ROC.

1 Introduction

Banks and financial institution provide a number of important financial services to businesses. One of the main services is granting credit to people and companies. For this, Credit risk has also been appeared as determinant of profitability of banks. The importance of credit risk evaluation was recognized since the 1988 Basel Capital Accord, which ignoring other kind of risks, set a minimal capital requirement for banks, based on their credit exposure.

In 1996 the Basel Committee proposed a variation to the original approach, taking into account also the market risk. The committee introduced the possibility to adopt an internal method to measure the market risk, through value at risk methodologies. The credit risk, instead, persisted to be treated using the standardized approach.

The new Basel Capital Accord of 2004, as well as recognizing the presence of operational risk, introduced, concerning credit risk quantification, the possibility to create an internal rating based system (IRB approach), where the banks compute only the default probability of their counterparts. The IRB approach requires therefore an internal rating system that, after being validated, would give banks the advantage to lower the capital requirement through a good credit policy.

Taking into account the Basel Committee's recommendations, it has become increasingly more important for banks to develop effective and reliable credit scoring systems to classify financially distressed firms. Several different methods and models have been utilized in doing so. Thomas (2000) briefly describes some of the techniques that have been used for credit scoring the last couple of decades. The most popular of these techniques is the frequentist logistic regression approach (Steenackers and Goovaerts (1989), (Laitinen, (1999), and (Alfo, Caiazza, and Trovato, (2005).

Although, it is important to have a credit scoring model with high predictive value, it is also important to account for random variation in the data. As the outcome variable is often a binary one representing whether or not the loan is granted, logistic regression model with random effects can be used. Wong and Mason (1985) originally proposed many applications of hierarchical logistic regression model. Avery et al. (2000) use logistic regression to study future loan performance and conclude that credit scoring improves the efficiency for the review process relative to solely depending on credit bureau scores.

However, some statisticians have recently argued that we stand at the threshold of a new Bayesian renaissance and other proponents argue that Bayesian methods more closely reflect how humans perceive their environment, respond to new information, and make decisions (Wylie, Muegge, and Thomas, 2006). In fact, Bayesian statistical analysis has benefited from the explosion of cheap and powerful desktop computing over the last two decades or so.

Bayesian methods are already increasingly being applied in a diverse assortment of fields, including medicine, sociology, psychology, artificial intelligence, and philosophy. It focuses on four essential elements. First, the incorporation of prior information is generally specified quantitatively in the form of a distribution and represents a probability distribution for a coefficient. Second, the combination of the prior with the likelihood function results in the creation of a posterior distribution of coefficient values. Third, simulates are drawn from the posterior distribution to create an empirical distribution of likely values for the population parameter. Fourth, basic statistics are used to summarize the empirical distribution of simulates from the posterior. Interested readers can consult a number of introductory texts focusing on the Bayesian perspective (Bolstad, (2004); Gelman, Carlin, Stern, and Rubin, (2004); Albert, J. (2007) ;Ntzoufras, I. (2009)).

Bayesian techniques have rarely been utilized by researchers or financial corporations in the past, but nowadays the increasing computational power entails that the computational challenges have been overcome (Wylie et al., 2006). In fact, Loffler et al. (2005) proposed a Bayesian method for banks to improve their credit scoring models by imposing prior information. This methodology enables banks with small data sets to improve their default probability estimates. Other authors like (Mira and Tenconi, 2003), Wilhelmsen et al. (2009), Fernandes et al. (2011) have already explored different Bayesian approaches for credit scoring, and found these methods to have some advantages over frequentist approaches.

In this paper we present the logistic regression model with random effects as a possible decision tool for credit scoring. Then, we explore Bayesian approach to estimating parameters models. The performance of Bayesian parameter estimation will be evaluated and compared with the parameters estimated by using a PQL method. Therefore, the research objective for this work is to analyze if Bayesian logistic regression is a more effective tool that improves quality of service and minimizes the risk of credit loss compared to a frequentist logistic regression.

The motivation in our study is to develop a Bayesian approach to estimate the logistic regression model with random effects. This paper will be divided into four sections. The first give a presentation of the data structure. The second section will be dedicated to explorer the PQL method and the Gibbs sampler algorithm. The third section will focus on empirical study to detect default Tunisian companies. Finally, the fourth will treat validation of the establish methods.

2 The data structure

2.1 The sample

The data used in this paper have been provided by the Central Bank of Tunisia. A series of financial data of 528 firms from different sectors (see Tab. 1) was collected from balance sheets and income statements for the period (1999- 2006). Our database is composed on 3065 files of credit that shows some heterogeneity due to business sectors diversity.

2.2 The explanatory variables

The financial ratios are usually used as predictors in failure prediction models; therefore, the choice of these independent variables is a fundamental problem. In our application, a battery of 26 ratios is used as inputs of the model which are defined in Tab. 2. These ratios are related to different dimensions of financial analysis and representing the different criteria for assessing the good health of company. Thus, themes are the financial structure,

	Sector	Number
1	Chemical Industry	34
2	Paper and paper board, publishing and printing	23
3	Extraction of non-energy	7
4	Transport and Communications	30
5	Agricultural and food industries	39
6	Manufacture of rubber and Plastics	27
7	Repair of motor vehicles and Trade household goods	69
8	Manufacture of machinery and equipment	26
9	Construction	36
10	Hotels et restaurants	37
11	Real estate renting and business services	23
12	Manufacture of leather and footwear	19
13	Agriculture, hunting, forestry	20
14	Textile and clothing	40
15	Manufacture of other non metallic mineral products	28
16	Metallurgy and Metalworking	27
17	Hospital and Health Social	21
18	Manufacturing and electronic equipment	13
19	Other manufacturing	20

Table 1: The number of firm by sector

$R_j(j1$ 13)	*Ratios definition*	$R_j(j1$ 23)	*Ratios definition*
R_1	Raw stock / Total assets	R_{14}	Rate of return on equity
R_2	Duration credit to the customer	R_{15}	Permanent capital turnover
R_3	Gross margin rate	R_{16}	Return on permanent capital
R_4	Operating margin rate	R_{17}	Rate of long-term debt
R_5	Ratio of personnel expenses	R_{18}	Ratio of financial independence
R_6	Net margin rate	R_{19}	Total debt ratio
R_7	Asset turnover	R_{20}	Immobilisation coverage by equity capital
R_8	Equity turnover	R_{21}	The long and medium term debt capacity
R_9	Economic profitability	R_{22}	Ratio of financial expenses
R_{10}	rate of return on assets	R_{23}	Financial expenses/total debt
R_{11}	Operating profitability of total assets	R_{24}	Working capital ratio
R_{12}	Gross economic profitability	R_{25}	Relative liquidity ratio
R_{13}	Net economic profitability	R_{26}	Quick ratio

Table 2: The inputs of model

rotation, profitability, financial expenses, solvency and liquidity.

2.3 The explained variable

The priori classification criterion adopted in this study is the state's legal business. Thus, the sample structure is described in two legal classes: healthy or default. The dependent variable can be written by binary values:

$$Y = \begin{cases} 1\ for\ default\ firm \\ 0\ for\ healthy\ firm \end{cases} \tag{1}$$

By adopting these criteria for classification, we could decompose a priori the sample into two subgroups. The first group is composed by 448 healthy firms and the second group is composed of 80 companies in distress.

3 The logistic regression model with random effects

3.1 Model Overview

Logistic regression is a probabilistic classification method where the probability of failure firm is estimating given its financial characteristics. It provides a linear function of the descriptors as a tool of discrimination. The study of this model is based on descriptors for binary variables and / or continuous variables. Indeed, logistic regression uses not only purely quantitative elements (the case of discriminant analysis), but it also incorporates qualitative factors (Bardos and Zhu, (1997)). Logistic regression is therefore of great interest.

Press and Wilson (1978) used annual cross section data ratios and countable sizes to examine whether the coefficients estimated of the logistic model are valid determinants of the firm's bankruptcy. Nevertheless, the significant information could be omitted by using only cross section analysis. In this case random effects models are frequently used to analyze complex data structures in the presence of significant sources of heterogeneity among individuals. Such models have been introduced in a wide variety of empirical applications, ranging from over dispersed to clustered observations. It has recently known a great interest due to the relevant impact of defaults credits on banks balances and to the proposal to modify the minimum regulatory capital by Basel Committee (2001).

3.2 Presentation of the econometric model

Logistic regression can be represented as an econometric method in which the endogenous variable Y is the encoding of companies: 0 if the firm has failed and 1 if the firm is healthy.

In this study, we selected 7 significant ratios. Given the structure of longitudinal data in our study, we applied the logistic regression model with random effects in the calculation of the risk of distress, taking into account the presence of a source of individual heterogeneity. The logistic regression model with random effects is written as follows:

$$\begin{aligned} log\left(\frac{P_{ij}}{1-P_{ij}}\right) &= \alpha + \beta_1 R_{7,ij} + \beta_2 R_{9,ij} + \beta_3 R_{10,ij} + \beta_4 R_{14,ij} \\ &\quad + \beta_5 R_{20,ij} + \beta_6 R_{21,ij} + \beta_7 R_{23,ij} + b_i, \end{aligned} \tag{2}$$

where R_k are financial ratio, β_k are unknown parameter and $p_{ij} = P(y = 1|X_{ij})$ with $i = 1, ..., 19 \quad and \quad j = 1, ..., n_i$ is the probability of belonging to the group of distress firms and b_i is the random effects which represent economic sector-specific random variation from the overall intercept, whose distribution is normal law $b \sim N(0, G_\delta$. This means that we associated a varying effect to intercept terms in each sector, to model heterogeneity present at sector level.

4 Estimation of logistic regression model with random effects

4.1 The penalized quasi likelihood method

The logistic distress scoring with random effect (2) can be estimated by the maximum likelihood analysis. The difficulties of determination explicit form of likelihood function has lead Breslow and Clayton, (1993) to develop new analytical approximation method and they give the name penalized quasi likelihood (PQL) method .

The technique PQL can estimate the parameters of the logistic regression model with random effects by adapting the problem to the estimation of the linear random effects model. In fact, the estimators of the model parameters by the PQL method are obtained by treating the random effects b as fixed parameters and the likelihood function is penalized according to the distribution of b. Thus, for a given value of θ, the estimators of the parameters β and b are obtained by maximizing the function of marginal log-likelihood penalized:

$$\log\{f(y|b)\} - \frac{1}{2}b^{'}G_{\theta}^{-1}b \tag{3}$$

The penalized marginal log-likelihood equation (3) is a non linear function with complicated shape. It is not possible to express estimators by simple observation functions. This equation must be solved by algorithms such as Newton-Raphson method which is based on the calculation of the first and second derivatives of equation (3). Breslow and Clayton (1993) have developed a formula similar to the Fisher scoring method of the linear random effects model.

Let $\mu = E(Y|X, Z, b)$ vector of the conditional mean of Y and $W = var(Y|X, Z, b)$ covariance matrix of Y, Differentiation of equation (3) with respect to β and b leads to the following normal equations:

$$g = \left[\begin{array}{c} X^{'}(Y-\mu) \\ Z^{'}(Y-\mu) - G_{\theta}^{-1}b \end{array} \right] \tag{4}$$

Whereas in the second order derivative of equation (3) with respect to β and b, give the following Hessian matrix:

$$H = - \left[\begin{array}{cc} X^{'}WX & X^{'}WZ \\ Z^{'}WX & Z^{'}WZ - G_{\theta}^{-1} \end{array} \right] \tag{5}$$

The parameters β and b of equation (2) can be determined iteratively by using the Newton-Raphson algorithm . Let $\delta = (\beta, \theta)$ a vector composed of the unknown parameters, at the iteration k, $\delta^{(k+1)}$ is calculated based on $\delta^{(k)}$ with the following recurrence formula:

$$\delta^{k+1} = \delta^{k} - \left\{H^{k}\right\}^{-1} g^{k} \tag{6}$$

Substituting equations (4) and (5) into equation (6), the following equations are obtained:

$$\left[\begin{array}{cc} X^{'}W^{k}X & X^{'}W^{k}Z \\ W^{k}X & Z^{'}W^{k}Z + W^{k}Z \end{array} \right] \left[\begin{array}{c} \beta^{k+1} \\ b^{k+1} \end{array} \right] = \left[\begin{array}{c} X^{'}W^{k}\tilde{y}^{k} \\ W^{k}\tilde{y}^{k} \end{array} \right] \tag{7}$$

where $\tilde{y}^{k} = X\beta^{k} + Zb^{k} + \left(W^{k}\right)^{-1}(Y - \mu^{k})$. Thus, using the pseudo data y_{pseudo}, parameter estimates (β, b) by the PQL method can establish

$$y_{pseudo} = X\beta + Zb + W^{-1}(y-\mu) = X\beta + Zb + \varepsilon_{pseudo}. \tag{8}$$

This equation has the same form of one of the random effects linear model, where W^{-1} is the inverse of the covariance matrix of pseudoerreurs ε_{pseudo}. According to the approach Breslow and Clayton (1993), the estimated logistic regression model with random effects (2) amounts to estimating a linear random effects model Indeed, by transforming the binary data to the explained variables as pseudo y_{pseudo} and calculating pseudo errors $\varepsilon_{pseudo} = W^{-1}(Y - \mu)$.

4.2 Gibbs Sampler algorithm

In this work, we choose to use a Bayes approach to the estimation of the parameters in logistic regression model with random effects (2). This approach requires the specification of prior distributions for σ^2 and regression coefficients in the β vector. The parameters are assumed to be randomly distributed across individuals, that is, for individual n , the vector of parameters β, follows a multivariate normal distribution $N(\mu,\Omega_\beta)$, where μ is the mean and Ω_β is the covariance matrix. We consider here a diffuse version of an inverse gamma distribution for the random effects variance σ^2.

The computational method often used is the Gibbs sampler, originally proposed by Geman and Geman (1984). An excellent discussion of this method can be found in Gelfand and Smith (1990). In order to obtain the Bayes parameter estimates, Markov chain Monte Carlo (MCMC) is used here. Although software such as the package (R2jags) of software R is now readily available for such Bayesian computations and the model proposed here are implemented, we include a short description of the method we use for simulating the posterior distributions of the model parameter estimates.

Bayes estimation procedures for the parameters in the model given by (2) require knowledge about the posterior distributions of these parameters. However, it is only possible to know these distributions up to a constant of proportionality; specifically, the posterior distribution for any given parameter is proportional to the product of all terms in the model that contain it. Therefore, for Model (2), if Y and b are vectors containing Y_{ij} and b_i respectively, while R is a matrix with rows R_{ij} , then

$$f(\beta_0|Y,\beta_1,...,\beta_7,b,\sigma^2,R) \propto \prod_{ij} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$$

$$f(\beta_1|Y,\beta_0,\beta_2,...,\beta_7,b,\sigma^2,R) \propto \prod_{ij} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}$$

$$f(b_i|Y,\beta,b_1,...,b_{i-1},b_{i+1},\sigma^2,R) \propto \prod_{ij} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} exp(-\frac{1}{2}\sum_i \frac{b_i^2}{\sigma^2})$$

$$f(\sigma^2|Y,\beta,R) \propto \frac{1}{\tau^{n+2}} exp(-\frac{1}{2}\sum_i \frac{b_i^2}{\sigma^2})$$

Under Gibbs sampling, an initial set of values are assumed as the estimates for β , b and σ^2 say $\hat{\beta}_{\{0\}}$, $\hat{b}_{\{0\}}$ and $\hat{\sigma}^2_{\{0\}}$. An updated estimate for β_0 , say $\beta_{0\{1\}}$ is obtained by sampling from the full conditional distribution $f(\beta_0|Y,\hat{\beta}_{1\{0\}},...,\hat{\beta}_{7\{0\}},\hat{b}_{\{0\}},\hat{\sigma}^2_{\{0\}},R)$ Sampling from the full conditional distribution$f(\beta_1|Y,\hat{\beta}_{0\{1\}},...,\hat{\beta}_{7\{0\}},\hat{b}_{\{0\}},\hat{\sigma}^2_{\{0\}},R)$ based on $\hat{\beta}_{0\{1\}}$ yields the revised estimate $\hat{\beta}_{11}$ for β_1. The completion of a first iteration is realized once the revised estimates $\hat{\beta}_{\{1\}}$, $\hat{b}_{\{1\}}$ and $\hat{\sigma}^2_{\{1\}}$ are obtained. This procedure of sampling using the most up-to-date revised estimates continues until the estimates of each parameter are deemed to have stabilized from one iteration to the next. See Geman and Geman (1984) and Gelfand and Smith (1990) for

	Estimes	P.discrim.	t-value	p-value
α	-2.258		-4.731	0.0000
R_7 Asset turnover	0.2744	0.0017	3.87	0.0001
R_9 Economic profitability	9.8965	0.5277	8.36	0.0000
R_{10} Rate of return on assets	-12.4456	0.4674	-8.40	0.0000
R_{14} Rate of return on equity	0.0327	0.0000	1.79	0.0740
R_{15} Capital turnover	-0.007	0.0011	-4.72	0.0000
R_{20} Immobilisation coverage by equity capital	-0.1935	0.0031	-4.82	0.0000
R_{21} The long and medium term debt capacity	-0.1341	0.0000	-2.64	0.0084
R_{23} Financial expenses/total debt	-0.8385	0.0000	-2.61	0.0091

Table 3: The estimated parameters of logistic regression model with PQL method

a discussion on Gibbs sampling, and Gelman and Rubin (1992) for methods of convergence.

5 The estimation results

5.1 The estimation results with PQL method

The logistic regression model with random effects was fitted on the available data using the package (glmmPQL) of software R. Six Fisher scoring iterations were needed for the algorithm, used to fit the model by the method of maximum marginal likelihood, to converge. The estimated parameters of the model are given in Table (3)

In the model (2) seven explanatory variables (R_7, R_9, R_{10}, R_{14}, R_{20}, R_{21} and R_{23}) have actually been selected among the ratios that have a major significance. These variables are significant at the 5% level of significance. This indicates that the variables included in the model are significant in explaining whether an applicant will be good or bad. The residual deviance of the model is 1,866.7 with 2,742 degrees of freedom.

	Mean .estime	sd.estime	2.5%	97.5%
α	-2.286	0.543	-3.395	-1.301
R_7 Asset turnover	0.241	0.107	0.035	0.451
R_9 Economic profitability	9.720	0.553	8.607	10.772
R_{10} Rate of return on assets	-12.166	0.627	-13.476	-10.969
R_{14} Rate of return on equity	0.039	0.019	0.004	0.079
R_{15} Capital turnover	-0.004	0.002	-0.009	0.000
R_{20} Immobilisation coverage by equity capital	-0.256	0.054	-0.368	-0.152
R_{21} The long and medium term debt capacity	-0.241	0.065	-0.376	-0.119
R_{23} Financial expenses/total debt	-0.529	0.300	-1.014	-0.099

Table 4: The estimates parameters of logistic regression model with Gibbs Sampler algorithm

The discriminant power of R_k is defined as the ratio : $\frac{\sigma_k^2\beta_k^2}{\sum \sigma_k^2\beta_k^2}$ with σ_k is the standard deviation of the ratio R_k. It expresses the influence of the ratio in the score function. According to Table (3) the ratios R_9 and R_{10} play a crucial role in the formation of the score function because these companies have a discriminating power ratios of around. In addition, we note that the estimated effect of the variable R_9 (economic profitability) has a positive sign. As the profitability ratio is the ratio between the financial costs and total assets. This means that the increase in financing costs reduced the profitability hence increasing the probability of being in distress. For the variable R_{10} (return on capital invested) that is equal to the ratio between net income and total assets has a negative sign which indicates that the increase in net results imply a reduction in risk of failure.

5.2 The estimation results with Gibbs Sampler Algorithm

Now, using the jags function from the package (R2jags) of software R, we fit logistic regression models with random effects. The Bayesian approach with informative priors use parameters from the standard logistic regression on the training data as priors. In order to obtain posterior estimates,we used a random walk Gibbs Sampler algorithm. Markov chain with 10,000 samples was generated for both models. The first 1,000 samples were excluded (to allow enough time for the Markov chain to converge to its stationary distribution) which left a Markov chain of 9,000 samples. Therefore, the burn-in period was 1,000.

The mean provides the estimate for the parameter. From Table (4), looking at the quantiles for each variable we can determine which variables are significant at the 5% significance level. The values from the 2.5% to the 97.5% quantiles provide a 95% credibility interval for each variable. All variables are significant. This shows that the majority of variables included in the model are significant in predicting good and bad applicants. The parameter estimates still have the same interpretation.

The estimate parameter of R_9 is 9.720 and is significant at the 5% significance level. The reason for this is that the 95% credibility interval does not contain zero. A unit increase in R_9 with all other variables held fixed, means that there will be 9.720 increase in the log-odds of default. For the variable R_{10}, the estimate parameter is -12.166 and is significant at the 5% significance level since its credibility interval does not contain zero. A unit increase in R_{10} with all other variables held fixed, means that there will be a 12.166 decrease in the log-odds of default.

The Figure (Fig .1) contain Trace plots of the Markov chain and density plots of the posterior distributions for the parameters of R_9 and R_{10} . From Figure 1, looking at the trace plot of the Markov chain, the Markov chain is relatively stationary. This implies that the Markov chain has reached or is close to its stationary distribution. A concern is that the Markov chain still appears to be quite strongly correlated.

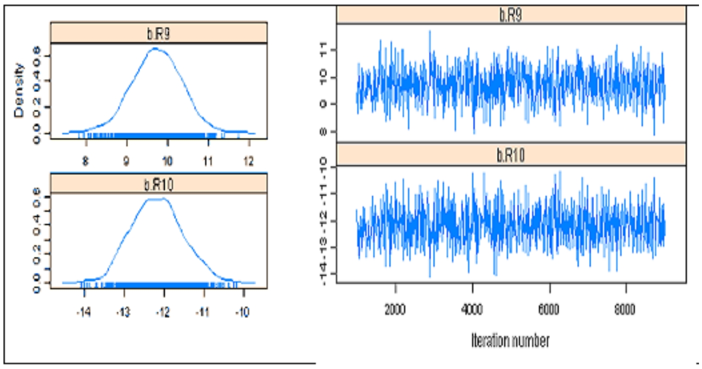
5.3 The estimation results of Random effects

After the integration of the sectoral effect in the logistic regression model, we produced the estimates presented in the Table (5). The result shows that the two approaches considered have very similar parameters estimates of the random effects. These estimates of random effects are the sectoral classification of sectors less risky to more risky. In other words from the

	Sector	random Effects with PQL method	random Effects with Gibbs Sampler
1	Repair of motor vehicles and Trade household goods	-4,401	- 4,825
2	Metallurgy and Metalworking	-2,943	-3,362
3	Manufacture of rubber and Plastics	-1,480	-1,562
4	Manufacture of leather and footwear	-1,009	-1,050
5	Agriculture, hunting, forestry	-0,768	-0,808
6	Manufacture of machinery and equipment	-0,654	-0,740
7	Hospital and Health Social	-0,596	-0,692
8	Real estate renting and business services	-0,256	-0,339
9	Manufacture of other non metallic mineral products	0,211	0,249
10	Textile and clothing	0,284	0,310
11	Chemical Industry	0,377	0,398
12	Transport and Communications	0,473	0,449
13	Manufacturing and electronic equipment	0,551	0,536
14	Extraction of non-energy	0,584	0,581
15	Paper and paper board, publishing and printing	0,597	0,623
16	Construction	0,860	0,895
17	Hotels et restaurants	1,045	1,012
18	Agricultural and food industries	1,198	1,201
19	Other manufacturing	6,261	6,712

Table 5: Estimated coefficients of random effects

Figure 1: Trace plots and density plots of the Markov chain for R_9 and R_{10}



1.PNG

	$\hat{Y}=1$	$\hat{Y}=0$	Total
$Y=1$	n_{11}	n_{10}	n_1
$Y=0$	n_{01}	n_{00}	n_0

Table 6: Confusion matrix

results of Table (5), the sector " Repair of motor vehicles and Trade household goods" is the least risky sector , since it has least value of random effect. Therefore, the sector "Other manufacturing" is the riskiest.

6 Validation of scoring functions of distress

After determining the score functions of distress by two different approaches, we must evaluate their effectiveness. We can do this by testing the discriminative power and predictive tests. Thus, we will calculate the rate of misclassification, plot the ROC "Receiver Operating characteristic"curve therefore calculate the area under curve (AUC) as a measure derived from the curve.

6.1 The rate of misclassification

To assess the ability to properly classify the model, we can construct a prediction column. Fixed a actually 0.5 as a cutoff, each firm is classified healthy if its probability of default is less than 0.5 and otherwise vulnerable. In practice, it is wiser to build what is called a confusion matrix (Table (6)).

	PQL method			Gibss Sampler		
	$\hat{Y}=1$	$\hat{Y}=0$	Total	$\hat{Y}=1$	$\hat{Y}=0$	Total
$Y=1$	8	30	38	23	5	28
$Y=0$	81	494	575	69	519	585
the rate of misclas.	0.181			0.115		

Table 7: Confusion matrix of the estimated models to the test sample

She always confronts the observed values of the dependent variable with those predicted, then records the good and bad predictions. The advantage of the confusion matrix is that it allows both to understand the error rate and realize the error structure.

The rate of misclassification is calculated by dividing the number of misclassification in the total sample $(n_{10}+n_{01}/n_0+n_1)$. According to Table (7), the rate of misclassification for the logistic regression model with random effect estimated by PQL method is equal to18.1% and 11.5% for the same model estimated by Gibbs Sampler. So we seen improved prediction of 6.6% using Gibbs Sampler algorithm. This proves the Bayesian approach overall outperforms the PQL method for determining risk of distress.

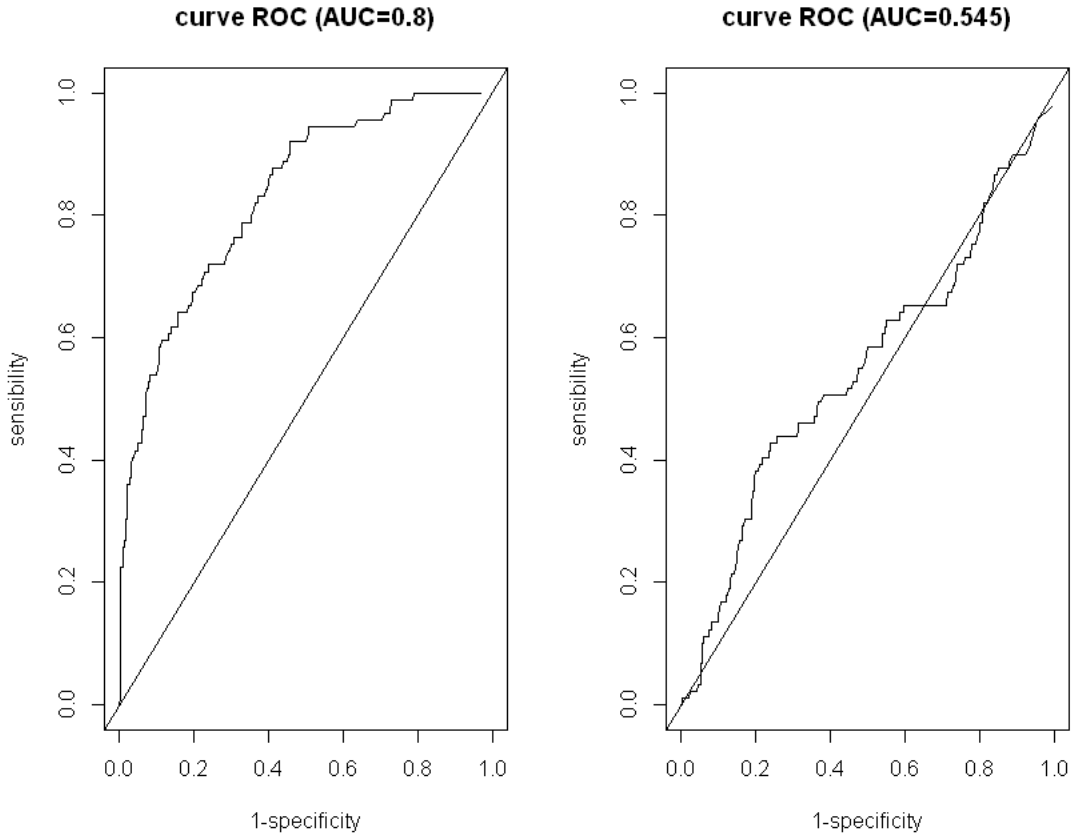
6.2 The curve ROC

Similarly, in order to compare PQL method and Gibbs Sampler algorithm for estimation of logistic regression model with random effects, we present the ROC curve of each method. This curve is a graphical tool to evaluate and compare the overall behaviour of the functions of scores (Pepe, 2000). The ROC curve relates the true positive rate (sensitivity) $(TPR = n_{00}/n_0)$ indicates that the model's ability to recover positives and false positive rate $(FPR = n_{10}/n_1)$ that corresponds to the proportion of negatives that were classified positive in a scatter graph. Usually, we compare with a threshold to make a prediction. We can build the matrix of confusion and extract the two indicators mentioned above. The ROC curve generalizes this idea by varying all possible values between 0 and 1. For each configuration, we construct the confusion matrix and calculate the (TPR) and (FPR).

In practice, it is not necessary to explicitly construct the matrix of confusion, we proceed as follows:

1. Calculate the score of each individual using the model prediction.
2. Sort the file by a decreasing score.
3. Consider that there is no tie. Each score value can potentially be a threshold s. For all observations whose score is greater than or equal to

Figure 2: The ROC curve of the model (2) estimated with PQL method and Gibbs Sampler



curve ROC (AUC=0.8)

curve ROC (AUC=0.545)

s, individuals in the upper table, we can count the number of positive and negative number.

4. The ROC curve is the graph that connects scatter the pairs (TPR, FPR). The first point is necessarily 1, the latter is 1. The procedure to calculate the cloud points of the ROC curve was performed using the software R.

According to the ROC curve (Fig .2) it is evident that the classification rule based on Bayesian logistic regression with random effects is more efficient than those based on standard logistic regression. This leads us to conclude that the Bayesian approach still perform better than PQL method for estimation the regression logistic model with random effects.

From this ROC curve, we can synthesize an indicator which reflects the predictive model. In fact, the AUC measures the quality of discrimination of the model and reflects the probability that a healthy company has a score above the score of a company in distress. The AUC of PQL method is equal to 0.540 whereas 0.792 for Gibbs sampler algorithm. These values are more close to one. This shows the advantage of using Bayesian approach and its impact on the predictive power of regression logistic model with random effects.

7 Conclusions

Credit Risk Management is assuming greater importance to all financial institutions. Thus, the risk prediction becomes an important issue. In this context several researchers have developed statistical tools to predict financial distress of companies.

This study provided an investigation into the use of Bayesian approach for estimation of logistic regression with random effects for credit scoring. The proposed technique presents various advantages. First the fact that the output of the Bayesian approach is the estimate of the posterior distribution of the default probability of each company. Having a distribution instead of a punctual value, we obtain a more complete and informative picture of the quantity of interest, that's to say the parameter uncertainty is also and easily taken into account during default prediction.

The second advantage is that, the logistic regression with random effects allows parametric flexibility among sectors to estimate default probability. They used to construct a sector classification based on the level of risk.

To compare the predictive performance of the Bayesian versus the classical model we performed a cross-validation analysis. We have compared the two different methods predictive ability (AUC) based on real data. By computing rate of misclassification classification for a fixed threshold, we show how the Bayesian approach overall outperforms the PQL method.

In conclusion, the results obtained show that the Bayesian approach is a powerful technique in terms of prediction relative to the PQL method. However, we can extend our research by using other Bayesian techniques such an algorithm Metropolis Hastings algorithm

References

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23.4,589-609.

[2] Avery, R.B., Bostic, R.W., Calem, P.S., and Canner G.B. (2000) Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files. *Real Estate Economics*, 28, 523-547.

[3] Albert, J. (2007). Bayesian Computation with R. New York: Springer Science+Business Media, LLC.

[4] Alfo, M., Caiazza, S. and Trovato, G. (2005),. Extending a Logistic Approach to Risk Modeling through Semiparametric Mixing. *Journal of Financial Services Research*, vol. 28, no. 1, pp. 163.

[5] Bolstad, W. M. (2004). Introduction to Bayesian statistics. Hoboken, NJ: John Wiley & Sons, Inc.

[7] Breslow, N. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal American Statistical Society* n 88, 9-25.

[8] Fernandes, G. and Rocha, C.A. (2011). Low Default Modelling: A Comparison of Techniques Based on a Real Brazilian Corporate Portfolio.

[9] Hand, D., and Henley, W. (1997), Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society A*, 160, 523-541.

[10] Mira, A. and Tenconi, P. (2004). Bayesian estimate of credit risk via MCMC with delayed rejection. In: Seminar on Stochastic Analysis, Random Fields and Applications IV. Centro Stefano Franscini, Ascona, pp. 277-291. Birkhauser Verlag, Basel

[11] Mestiri, S., Hamdi, M.(2012). Credit Risk Prediction: A Comparative Study Between Logistic Regression and Logistic Regression with Random Effects *International Journal of Management Science and Engineering Management* 7 (3), Taylor & Francis, 200-204.

[12] Hamdi, M., Mestiri, S. (2014). Bankruptcy Prediction For Tunisian Firms: An Application Of Semi-Parametric Logistic Regression and Neural Networks Approach. *Economics Bulletin* 34 (1), AccessEcon, 133-143.

[13] Laitinen, E.K. (1999) Predicting a Corporate Credit Analyst's Risk Estimate by Logistic and Linear Models. *International Review of Financial Analysis*, vol. 8, no. 2, pp. 97.

[14] Steenackers, A. and Goovaerts, M.J. (1989) A Credit Scoring Model for Personal Loans. *Insurance Mathematics and Economics*, vol. 8, no. 1, pp. 31.

[15] Loffler, G., Posch, P.N., and Schone, C. (2005). Bayesian methods for improving credit scoring models. Technical report, Department of Finance, University of Ulm, Germany.

[16] Ntzoufras, I. (2009). Bayesian Modeling Using WinBUGS. John Wiley & Sons, Inc., Hoboken, New Jersey.

[17] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian Data Analysis (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

[18] Geman, S., and Geman, D. (1984), Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[19] Gelfand, A.E., and Smith, A.F.M. (1990), Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398-409.

[18] Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association,* 95 :308-311.

[19] Press, S. J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association,*73:699-705.

[20]Thomas, L.C. (2000), A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. *International Journal of Forecasting*, vol. 16, no. 2, pp.149.

[21] Wong, G.Y., and Mason, W.M. (1985), The Hierarchical Logistic Regression Model for Multilevel Analysis. Journal of the American Statistical Association, 80, 513-524.

[22] Wylie, J., Muegge, S. and Thomas, D.R. (2006), Bayesian Methods in Management Research: an Application to Logistic Regression

[23] Wilhelmsen, M., Dimakos, X.K., Huseb, T., and Fiskaaen, M. (2009). Bayesian Modelling of Credit Risk using Integrated Nested Laplace Approximations. Available from http://publications.nr.no BayesianCreditRiskUsingINLA.pdf.

[24] Ziemba, A. (2005). Bayesian Updating of Generic Scoring Models. Available from http://www.crc.man.ed.ac.uk/conference/archive/2005/papers/ziemba-arkadius.pdf. Accessed date: 25th November 2010.