ACTA TECHNICA CORVINIENSIS – Bulletin of Engineering
Tome V (Year 2012). FASCICULE 4 [October–December]. ISSN 2067-3809

[1.] Snežana KONJIKUŠIĆ, [2.] Lidija BARJAKTAROVIĆ, [3.] Ana VJETROV

APPLYING OF STATISTICAL MODEL IN DETERMINING THE MINIMAL NUMBER OF OBSERVATIONS FOR CALCULATING CREDIT DEFAULT RATE

[1,3.] FACULTY OF ECONOMICS FINANCE AND ADMINISTRATION (FEFA), SERBIA
[2.] SINGIDUNUM UNIVERSITY, SERBIA

ABSTRACT: *The important aspect related to both banks' and other financial institutions' performances is the analysis of the clients' obligations fulfillment through applying various mathematical models. Furthermore, the main purpose of the models is estimation of the possibility of default (default rate). Moreover, it is important for the users of the model to posses data related to possibilities' accuracy. There are various ways for testing the accuracy related to the default possibilities, and in this paper the minimal number of observations will be presented. Thus, simple statistical operations will be presented, important for gaining a minimum level related to sample size. Moreover, the paper will point out that for determining the minimum size of the sample, the absence of correlation among data is essential. If the correlation is present, the minimum level of accuracy of the sample size could be altered, representing the important conclusion of this paper. Furthermore, if there is fixed sample, this approach allows minimum difference between estimated and actual empirical default rate. Finally, the cases indicating inaccurate minimum level of the sample size will be shown as well.*
KEYWORDS: *default probability, default rate, default rate prediction, minimum level of sample size*

INTRODUCTION

It is a common thing for credit model users to calculate default probability through credit rating of the clients. Moreover, they are inclined to know the precise probabilities related to defaults. Consequently, the experiments estimating the difference between expected and actual bankruptcy rate are frequently undertaken (default rates).

However, there are various methods for the calculation. These cases imply further analyses related to bankruptcy rates' fitting within expected range of credit rating (estimating the difference between the expected and actual bankruptcy rate). Usually, these cases involve large samples, especially when the probabilities are low, as it is the case with high credit rating. Nevertheless, how big the sample should be is the frequently asked question. Thus, the aim of the paper is to calculate the number of observations necessary for the tests.

The approach includes statistical relations. Consequently, the results point out that the lowest level of the sample size can be implemented only when there is an absence of correlation among data. In other words, the lower limit can be calculated, when there is no correlation between time and cross sectional data.

Furthermore, the values of correlations influence the value related to lower level of sample size as well. For example, when the correlation coefficient is zero and the sample size is large enough, the lowest level can be small enough. However, when correlation coefficient differs from zero, the adding of observation units would not contribute to the narrowing of confidence interval.

Particularly, the analytical link can be suitable for decision making related to the size number of available data and for probabilities estimation. Otherwise, when the fixed size sample is used, this can be helpful in determining the minimum difference between the expected and actual size of default rate, which is statistically significant. Furthermore, within the cases where the initial hypotheses are jeopardized, the using of simulation model is highly recommended.

THEORETICAL BACKGROUND

Theoretical background of the paper is related to Law of Large Number and the Central Limit Theorem. Furthermore, binomial distribution is used and its tendency to have characteristics related to normal distribution when it is a case with large number of observations. The latter includes situations when the data are independent.

Binomial distribution is determined by two parameters p and n, representing default probability and numbers of observations, i.e. number of companies. Nevertheless, it is essential to stress out that out of n number of companies d number present default companies. Consequently, the aim is to determine whether expected default rate is near the level of actual one. Using relative frequency, default rate is defined as:

$$f_d = d/n \qquad (1)$$

Under the following condition:

$$P(|f_d - p| < \varepsilon) \leq \alpha \qquad (2)$$

where α present stands for the level of significance. Assuming that "default rate" is binomial random variable, there are two cases: default or not. Binomial distribution for large enough number of observations (n) converges to normal distribution. Using CLT the result is:

$$P(np_L \le np \le np_U) \approx \frac{1}{\sqrt{2\pi}} \int_{\frac{n(p_L-p)}{\sqrt{npq}}}^{\frac{n(p_U-p)}{\sqrt{npq}}} e^{-x^2/2} dx =$$

$$= \Phi\left(\frac{n(p_U-p)}{\sqrt{npq}}\right) - \Phi\left(\frac{n(p_L-p)}{\sqrt{npq}}\right) \quad (3)$$

Φ representing density function related to normal distribution. Assuming that:

$$p_U\text{-}p = p\text{-}p_L = \varepsilon .$$

It is easily noticeable that $(p\text{-}\varepsilon, p+\varepsilon)$ presents confidence interval default rate at α significance level. Using relation (3), the following results are acquired:

$$2\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) - 1 \ge 1 - \alpha$$

that is

$$\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) \ge 1 - \frac{\alpha}{2}$$

$$\frac{n\varepsilon}{\sqrt{npq}} \ge \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (4)$$

$$n \ge \frac{pq}{\varepsilon^2}\left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]^2$$

The last row provides the formula essential for calculation of n, representing a minimal number of observations or companies, following default probability p, level of accuracy ε, and level of significance α.

Moreover, in cases including n observations, in order to determine the difference between p and f_d, under the previously set level of significance, the following inequality is used:

$$\varepsilon \ge \sqrt{\frac{pq}{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

In conclusion, under the expected probability, the further analysis estimates its suitability for calculation of default rate. On the other hand, the question related to the choice of number of the companies necessary for the analysis stays open. Furthermore, it is observed that the figure pq reaches its maximum level providing that p=q=0.5. By setting p=0.5, the calculation of default rate is possible with accuracy, providing there is a minimum number of the companies. Moreover, when the number of the companies is fixed, default rate estimation number will be within ε with the probability of $100*(1-\alpha)\%$. Additionally, this is a standard confidence interval for probability estimation.

Nevertheless, the samples used are significantly smaller than the population, comprising 5-10% out of the overall population. In cases when population is finite, correction factor (N-n)/(N-1) is used, where N represents number of population members, and n presents number of sample members.

The following table presents the values for ε, with the number of observations (companies) n, p=0.005 and α=0.05. Consequently, the table shows that the increase of number of observations is followed by the change related to the value of ε.

Table 1. The values of ε when n, p and α are familiar

n	ε
1000	0.0044
2000	0.0028
5000	0.0020
10000	0.0014

METHODOLOGY

Additionally, it is very useful to analyse n/p, when they present small values. Consequently, it is not recommendable to use analytical results. On the other hand, simulation is more complicated but trusting mechanism determining right values.

The following table shows examples related to simulation and analytical results for the value ε, while n, p and α are familiar.

Table 2. Analytic vs. Simulated levels of ε

n	P	Analytic ε	Simulated ε	Percent difference
100	0.001	0.0062	0.0090	-23%
250	0.001	0.0039	0.0030	8%
500	0.001	0.0028	0.0030	2%
100	0.025	0.0306	0.0250	-18%
250	0.025	0.0194	0.0190	-2%
500	0.025	0.0137	0.0130	-5%

The previous table indicates that analytical result enable acceptable estimation in cases when n/p has higher values. Otherwise, relative difference among predicted results (1-ε analytical/ε simulation) may appear significantly high, in cases concerning small values. The result acquired recommends avoiding approximation when npq is less than 2. Furthermore, these cases include possibility of asymmetrical distribution, thus complicating the interpretation of the results (asymmetry of binomial distribution is presented as 1-6pq/npq). Informal experiments recommend simulation in cases when npq is less than 4. Furthermore, the experiments include relative errors less than 10%, predicting ε.

Unfortunately, the latter does not present the best mechanism for estimation of n using simulation. However, analysts more frequently have fixed samples of the companies than they have fixed values for ε, making it not as practical problem as it could be. Furthermore, it is quite feasible to use methodology for calculating specific values for n, by using inequality (4) already presented in the paper. Consequently, the approximation fit well, except for small values. Excluding extreme values, errors were below the level of 10%.

Hence, the previous section deals with setting the low limit. Statistical theory, applying CLT restrict both values and the upper limits. However, it is seldom case that companies' databases fit conditions imposed by CLT. Thus, hypotheses annulling may appear as the result of additional variances and covariance affecting the higher values for n and ε.

Additionally, the companies' rating may be determined by estimated probability. Moreover, the estimated values may cause the increase of sample's variability.
Furthermore, the analysis presented on the following two tables clearly indicates that the result appear to be worse when there is a correlation among data. Hence, it is advisable to avoid analysis related to the same company in certain period of time, influenced by the same economic factor.
Table 3 shows values related to ε when n, p, correlation coefficient are already given and α =0.05.

Table 3. Values for ε when n, p, α are familiar

Correlation	N	p=0.01	p=0.03	p=0.05
0.0	500	0.008	0.011	0.018
0.1	500	0.020	0.048	0.07
0.2	500	0.030	0.063	0.108
0.3	500	0.036	0.083	0.142
0.1	1000	0.006	0.008	0.011
0.2	1000	0.020	0.046	0.067
0.3	1000	0.029	0.061	0.102

Thus, it is clearly noticeable that ε has the smallest value when correlation coefficient equals zero, which presents the main goal, due to the fact that ε presents difference between actual and expected probability at given level of significance.
Furthermore, the following table shows values for ε when probability has only one value, i.e. 0.01, level of significance is 0.05, correlation coefficient has only two values and finally and number of observations lies within the range from 100 to 1000.

Table 4 .Values for ε when n, p, α are familiar

Correlation	n	p=0.01
0.03	100	0.040
0.03	250	0.038
0.03	500	0.036
0.03	1000	0.034
0.00	100	0.020
0.00	250	0.010
0.00	500	0.008
0.00	1000	0.006

Having presented the results, it is evident that the accuracy of probability is more influenced by the independence of the data than by the number of observations. Additionally, the independence appears to be very important due to convergence of binomial distribution towards normal.
Thus, it is essential to mention that the increase of correlation coefficient is followed by the increase of distribution asymmetry. However, the biggest weakness related to parameters' asymmetry is their interpretation. Nevertheless, table 4 clearly indicates that there is asymmetrical distribution for small values of n as well (values smaller than 500).
Moreover, table 4 indicates that distribution becomes asymmetrical when correlation coefficient equals zero and n value is low. Consequently, as long as n is less than 500, both theoretical and simulated predictions for ε are lower than p.
However, it is difficult to estimate correlation coefficient in practice, due to samples' overlapping. Thus, plenty of companies would be included in the sample for more than a year, causing overlapping. Furthermore, default rates could not be constant within time. Finally, various correlation effects influence the change of default rate' variance.
To summarize, there is a great deal of evidence for actual implementation of predicted values of p and n using inequalities (2) and (3), ignoring unexpected external effects.

CONCLUSIONS

In conclusion, while using mathematical model for calculating defaults' probabilities, it is essential to determine the probabilities' accuracy as well. Consequently, the estimation of actual and predicted default rates has to be done. Furthermore, statistical method is implemented in order to determine the lowest limit or number of observations needed accurately.
The model's presence, setting the lowest limit is useful when rigorous conclusions related to probabilities' estimation are not made. Furthermore, determining the lowest number of observation does not include positive correlation among data.
On the other hand, if the sample is fixed, this approach may be used to set the difference between actual and predicted default rate.
Finally, as long as the values for ε and n are lower and do not satisfy pointed inequalities, the result obtained would not be statistically significant.

ACKNOWLEDGEMENTS

This Research Paper was the part of the project "Improvement of Serbian competitiveness in the process of entering to European Union", no. 47028, in the period 2011-2015, financed by Serbian Ministry of science and technological development.

REFERENCES

[1.] Bangia, A.,F.X. Diebold, A. Kronmus, C. Schagen and T. Schuermann. Ratings Migration and the Business Cycle, With Applications to Credit Portfolio Stress Testing. Journal of Banking &Finance 2002 26:445/474
[2.] Barjaktarović, L. [2009] Risk management. Singidunum University
[3.] Barjaktarović, L. and Ječmenica, D.[2011]. Optimism vs. pessimism of competitiveness of Serbian banking sector. Industrija no 2/11, p 137-150
[4.] Barjaktarović, L., Popovčić-Avrić, S. I Đenić, M. [2011] Collection management as curtail part of credit risk management during the crisis, 8th AFE 2011 Conference Samos, Greece Cantor, R. and F. Packer.
[5.] Christensen, J.E. Hansen and Lando,D. [2004]. Confidence Sets for Continuous-Time Rating Transition Probabilities. Journal of Banking & Finance 28:2575-2602.
[6.] Gagliardini, P. and Gourieroux, C. [2005] Stochastic Mogration Models with Application to Corporate Risk. Journal of Financial Econometrics 3:188-226.
[7.] Hanson, S.G. and Schuermann,T. [2006] Confidence Intervals for Probabilities of Default. Journal of Banking & Finance 30:2281-2301.

[8.] Jafry, Y. and T. Schuermann. Measurement, Estimation and Comparison of Credit Mogration Matrices. Journal of Banking & Finance 200428:2603-2639.
[9.] Konjikusic, S. Barjaktarovic, L. and Radojevic, G. Methodology of calculating probability of default, Market money capital 2011 4:31-40
[10.] Lando, D. and T. Skodeberg. Analyzing Ratings Transitions and Rating Drift with Continous Observations. Journal of Banking&Finance 2002 26:423-444
[11.] Mahlmann, T. [2005]. Biases in Estimating Bank Loan Default Probabilities. Journal of Risk 7:75-102.
[12.] Pluto, K. and D.Tasche. Estimating Probabilities of Default for Low Default Portfolios Risk 2005 18:76-82
[13.] Radojević, G. i Konjikušić, S. [2004]. Investment choice using Promethee methods. SymOpis
[14.] The Credit Rating Industry. Journal of Fixed Income, 1995 December: 10/34
[15.] Truck, S. And Rachev, S. [2005]. Credit Portfolio Risk and Probability of Default Confidence Sets through the business Cycle. Journal of Credit Risk 1:61-88.

ACTA TECHNICA CORVINIENSIS – BULLETIN of ENGINEERING

ISSN: 2067-3809 [CD-Rom, online]