FINANCE MANAGEMENT AND BANKING

DEVELOPMENT OF DEFAULT MODELS UNDER LIMITED DATA ACCESS CONDITIONS

Jānis Bokāns, Dr.paed.
DnB NORD Bank,
E-mail: janis.bokans@dnbnord.lv

Marina Kudinska, Dr.oec.
Latvian University
E-mail: marinak@lu.lv

Irina Genriha, Mg.oec.
DnB NORD Bank,
E-mail: irina.genriha@dnbnord.lv

Abstract

Since 2008 in banking sector capital requirements are set by "Basel II" approach. The internal rating models (IRB) capital calculation approach is based on probability of default (PD) models.

Historical data availability to develop PD statistical models often is limited. The model development must take into account the statistical regularities, including possible model overloading.

To develop probability of default statistical model for small and medium enterprises the modelling data set from the 2800 financial statements was used. The data set include 54 events of default.

The 34 debt coverage, liquidity, profitability, activity and other financial indicators were examined. Model overloading effects on the statistical reliability was investigated. By various statistical tests and methods the optimal number of risk factors was established and several models with different number of risk factors were compared.

It was shown that the model with optimal number of risk factors demonstrated the best statistical reliability. The results were analyzed in relation to minimum sample size criteria available in literature.

Keywords: probability of default, risk capital, statistical models, sample size, power

1. Introduction

Following the Revised International Capital Framework by the Basel Committee on Banking Supervision (known as Basel II), qualified banks is going to use the Internal Rating-Based (IRB) approach for economical capital calculation. One of the important IRB components is the measuring the credit risk by the Probability of Default (PD), which are to be estimated by banks applying their internal credit risk models. Quality of internal credit risk models is therefore of key importance for the calculation of economical capital. For probability of default model development historical data are required. This paper offers some suggestions for determination of sample size providing pre-defined preciseness power for the probability of default model.

2. Sample size

The traditional approach to statistical model building involves seeking the most parsimonious model that still explains the observed data. The rationale for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and is more generalized.

Sample size determination is often an important step in planning a statistical study and it is usually a difficult one.

One could suggest two important questions regarding sample size:

1. How many subjects do we need to statistically prove observed influence of specific factor?
2. Do we have enough data to fit the model?

Whitemore research results provide some guidance for a logistic model containing a single dichotomous covariate (cited from Hosmer & Lemeshow, 2000). On default model example we illustrate one of the Whitemore (1981) approaches. We use it to evaluate what sample size we would take to test statistical reliability of 50 percent increase in the default frequency.

In terms of the logistic regression model the null and alternative hypotheses are $H_0 : \beta_1 = \ln(1) = 0$ versus $H_a : \beta_1 = \ln(1.5)$. To determine the sample size we need an estimate of the response probability $P_o = (Y = 1 | x = o)$. Cross classifying the outcome variable (PD) by the covariate (for example: Capital Ratio) shows that 20 percent of observations with Capital ratio > 0.5 are default. In this case the formula is

$$n = (1 + 2P_0) \times \frac{\left(z_{1-\alpha} \sqrt{\frac{1}{1-\pi} + \frac{1}{\pi}} + z_{1-\theta} \sqrt{\frac{1}{1-\pi} + \frac{1}{\pi e^{\beta_1}}} \right)^2}{P_0 \beta_1^2}$$

where $z_{1-\alpha}$ and $z_{1-\theta}$ denote the upper α and θ percent point respectively of the standard normal distribution. This number we would need for a 5 percent level test to have 80 percent power.

π- denotes the fraction of subjects in the study expected to have x=0. In our case we have unequal numbers of events in covariate. Let the value $\pi = 0.9$. The sample size is

$$n = (1 + 2 \times 0.2) \times \frac{\left(1.645 \sqrt{\frac{1}{1-0.9} + \frac{1}{0.9}} + 0.842 \sqrt{\frac{1}{1-0.9} + \frac{1}{0.9 e^{[\ln(1.5)]}}} \right)^2}{0.2 \times [\ln(1.5)]^2} = 2892$$

This denotes that, rounding up, we would need 2892 subjects, from which 10 percent is default cases.

A second consideration, and one relevant to any model being fit, is the issue of events per covariate. Peduzzi, Concato, Kemper, Holford and Feinstein (1996) examine the issue of how many events per covariate are needed to obtain reliable estimates of regression coefficients when fitting a logistic regression model. In general the relevant quantity is the frequency of the least frequent outcome, m=min (n1, n0). In our case this is usually the number with the event present (y=1) but it could just as well be the number with the event absent (y=0). Peduzzi et.al. show that a minimum of 10 events per covariate are needed to avoid problems of over estimated and under estimated variances and thus poor coverage of Wald-based confidence intervals and Wald tests of coefficients.

Thus the simplest answer to the „do we have enough data" question is to suggest that the model contain no more than $p + 1 \le \min(n_1, n_0)/10$ covariates.

In our case we have 54 defaults and 2746 non-defaults, what mean the follows: the rule suggests that the models should contain no more than 4 covariates.

$$p + 1 \le \frac{\min(54; 2746)}{10} = 5$$

Green (1991) provides a comprehensive overview of the procedures used to determine regression sample sizes. Although there are more complex formulae, the general rule of thumb *is no less than 50 subjects for a correlation or regression modelling with the number increasing with larger numbers of independent variables.* Green suggests $N > 50 + 8$ m (where m is the number of independent variables) for testing the multiple correlation and $N > 104 + m$ for testing individual predictors. If testing both, use the larger sample size.

Although Green's (1991) formula is more comprehensive, there are two other rules of thumb that could be used. *With five or fewer covariates* (this number would include correlations), a researcher can use Harris's (1985) formula for yielding the absolute minimum number of subjects. Harris suggests that the number of subjects should exceed the number of covariates by at least 50 (i.e., total number of subjects equals the number of covariate variables plus 50). Larger samples are needed when the depended variable is skewed, the

effect size expected is small, there is substantial measurement error, or stepwise regression is being used (Tabachnick & Fidell, 1996).

Based on the above mentioned studies, we decided to examine and verify in practice these assumptions. As a result we opt for the assumptions Peduzzi, Concato, Kemper, Holford and Feinstein (1996) and portrayed our results in the Figure 1.

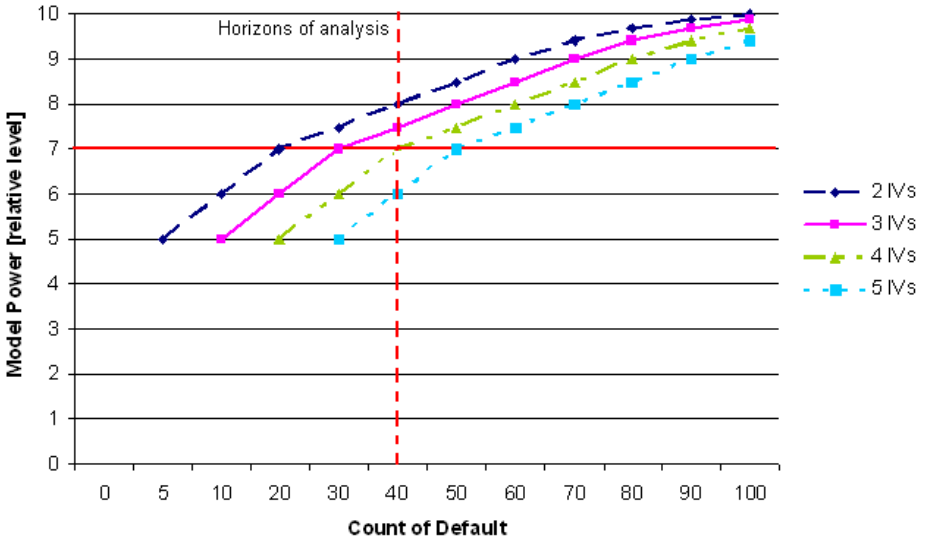

Figure 1: Count of Default vs Model Power and Count of Independent Variables in the model

Following our analysis we concluded that Peduzzi, Concato, Kemper, Holford and Feinstein (1996) assumption is the more appropriate in practice to evaluate the necessary size of historical data sample. Based on this approach and our historical data base we analyzed and developed a graphic model with which is possible to determine: how much count of default need to have to develop a model with specific degree of power or how many factors in the model we can use in a certain sample size or if we have certain count of default and count of covariate how high power is possible to achieve. At example if we have 40 default cases and we have 4 covariates it is possible to achieve only level 7 of model power [1], but if we want to achieve a possible maximum level of model power it's necessary to use only 2 covariates. Of course if we have not sufficient count of defaults it is not possible to achieve 10 level of model power, because based on statistical practice experience the minimum count of default events is 100, but unfortunately, in practice we often meet with cases where the sample size is very small.

3. Variable selection in the Binomial Logistic Regression

The criteria for including a variable in the model may vary from one problem to another and from one scientific discipline to another. The more variables are included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on specific historical data set. Researcher suggest including all practically and intuitively relevant variables in the model regardless of their statistical significance. The major problem is that the model may be „over fit", producing numerically unstable estimates. Over fitting is typically characterized by unrealistically large estimated coefficients and/or estimated standard errors. This may be especially troublesome in problems where the number of

1 Model power relative level implies a combined level of three statistical tests: ROC, Kolmogorov-Smirnov and Hosmer-Lemshow

variables in the model is large relative to the number of subjects and/or when the event is close to either 0 or 1.

In variable selection process it is important to pay attention to number of subjects per each variable to see how clustering affects the tests of independence (chi-square). The chi-square test of independence is used to determine whether there is a relationship between two categorical variables. The chi-square statistic measures the overall discrepancy between the observed cell counts and the counts you would expect if the column proportions were the same across columns. A larger chi-square statistic indicates a greater discrepancy between the observed and expected cell counts—greater evidence that the column proportions are not equal, meaning hypothesis of independence is incorrect.

In the cross table more than 20% of each table's cells have expected counts no less than 5, and the minimum expected cell count is no less than 1. These notes indicate that the assumptions of the chi-square test may not be met by these tables, and so the results of the tests are suspect. (SPSS Manual)

However, the number of subjects still impacts the power. Small expected frequencies in one or more cells limit power considerably.

All difference-detecting tests are based on covariate distribution in the sample. The more observations, the narrower the distribution, and the greater the likelihood, that any differences will be discovered (i.e., the greater the power.) Power is not, however, only related to sample size it is also related to effect size. The greater the effect size is, the greater the power.

For example, the difference in default frequency between new enterprises until 1 year old and young enterprises until 3 year old is likely very small; therefore, the effect size is small. The difference in default frequency between enterprises until 1 year old and 5-10 years old enterprises is much larger; hence there is a larger effect size and a greater ability to detect differences (greater power).

There are several steps one can follow to aid in the selection of variables for a logistic regression model. The selection process begin with a univariate analysis of each variable (potential input), in order to select the most intuitive and powerful variables. This analyse is providing by the graphical description of the relationship between each variable individually and the default frequency. To avoid *multicollinearity problem*, the explanatory variables of regression should be not correlated between themselves, because inclusion of highly correlated variables for estimation of the optimal weights for a model can result in unstable estimates of those.

Binomial Logistic Regression is a type of regression useful to model relationship where the dependent variable is dichotomous (only assumes two values) and independent variables are of any type. Logistic regression estimates the probability of a certain event accruing, since it applies maximum likelihood estimation after transforming the dependent variable into a logit variable (the natural log of the odds of the dependent occurring or not).

The more important task for modelling decision then functional form of model is the transformation of independent variables. For this purpose we use the following logistic function:

$$L(x;a,b)=\frac{1}{1+\exp\{-(a+b\cdot x)\}},$$

where a, b are constants.

The last step to variable selection is to use a stepwise method in which variables are selected either for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. There are two main versions of the stepwise procedure: forward selection with test for backward elimination and backward elimination followed by a test for forward selection. This procedure is available in SPSS program.

4. Model validation

A model with high discriminatory power is a model that can clearly distinguish the default and non-default populations. In other words, it is a model that makes consistently "good" predictions relative to few "bad" predictions. For a given cut-off value (PD threshold), there are two types of "good" and "bad" predictions:

Table 1: Cross table

		Estimated	
		Non-Default	Default
Observed	Non-Default	True	False Alarm (Type II Error)
	Default	Miss (Type I Error)	Hit

The "good" predictions occur if, for a given cut-off point, the model predicts a default and the firm does actually default (Hit), or, if the model predicts a non-default and the firm does not default in the subsequent period (True).

The "bad" prediction occurs if, for a given cut-off point, the model predicts a default and the firm does not actually defaults (False-Alarm or Type II Error), or if the model predicts a non-default and the firm actually defaults (Miss or Type I Error).

The Hit Ratio (HR) corresponds to the percentage of defaults from the total default populations that are correctly predicted by model, for a given cut-off point.

The False Alarm Ratio (FAR) is the percentage of False Alarms or incorrect default predictions from the total non-defaulting population, for a given cut-off point.

Several alternatives could have been considered in order to analyze the discriminating power of the estimated model. In this study, both ROC/CAP analysis and Kolmogorov-Smirnov (KS) analysis were performed.

Receiver Operating Characteristics (ROC) curve is a plot of the HR against FAR for all possible cut-off points

Cumulative Accuracy Profiles (CAP) curve is a plot of the HR against the percentage volume of the sample.

The Kolmogorov-Smirnov (KS) methodology considers the distance between the distributions of 1-HR (or Type I Errors) and 1-FAR (or True predictions). The higher distance between the two distributions, the better the discriminating power of the model.

For the ROC curve, a perfect model would pass through the point (0.1) since it always makes „good" predictions, and never „bad" predictions (it has FAR = 0% and a HR=100% for all possible cut-off points). A „naive " model is not able to distinguish defaulting from non-defaulting firms, thus will do as many „good" as „bad" predictions, though for each cut-off point, the HR will be equal to the FAr. A better model would have a steeper curve, closer to the perfect model, thus a global measure of the discriminant power of the model would be the area under the ROC curve. This can be calculated has:

$$AUROC = \int_0^1 HR(FAR)d(FAR)$$

For the CAP, a perfect model would attribute the lowest scores to all the defaulting firms, so if x% of the total population are defaults, then CAP curve of a perfect model would pass through the point (x, 1). A random model would make as many „good" as „bad" predictions,

so for the y% lowest scored firms it would have a HR of y%. Then, a global measure of the discriminant power of the model, the Accuracy Ratio (AR), compares the area between the CAP curve of the model being tested and the CAP of the random model, against the area between the CAP curve of the perfect model and the CAP curve of the random model.

It can be shown that there is a linear relationship between the global measures resulting from the ROC and CAP curves:

$$AR=2*(AUROC-0.5)$$

AR value is depending from default frequency in the modelling sample size too, the smaller the default frequency, the more sensitive model assessment and ROC/CAP curves and the greater AR value it is possible to achieve (look Table 2).

Table 2: ROC area value vs Default Frequency

Default count / Total sample size	Default count	Default Frequency	Max AR
1/1000	1	0.1%	99.9%
10/1000	10	1.0%	99.5%
20/1000	20	2.0%	99.0%
30/1000	30	3.0%	98.5%
40/1000	40	4.0%	98.0%
50/1000	50	5.0%	97.5%
100/1000	100	10.0%	95.0%
150/1000	150	15.0%	92.5%
200/1000	200	20.0%	90.0%
250/1000	250	25.0%	87.5%
300/1000	300	30.0%	85.0%
400/1000	400	40.0%	80.0%
500/1000	500	50.0%	75.0%

In order to test the overall significance of the model, the **Hosmer-Lemeshow test** was used. This "goodness-of-fit" test compares the predicted outcomes of the logistic regression with the observed data by grouping observations into risk deciles (g=10) and then computing a Pearson chi-square statistic that compares the predicted to the observed frequencies in a 2x10 contingency table. Let o_i^0 be the observed count of non-defaults for group i and p_i^0 be the predicted count. Similarly, let o_i^1 be the observed count of defaults for group i and p_i^1 be the predicted count. Then the HL test statistic following a chi-square distribution with g-2 degrees of freedom is:

$$HL=\sum_{i=1}^{g}\left[\frac{(o_i^0-p_i^0)^2}{p_i^0}+\frac{(o_i^1-p_i^1)^2}{p_i^1}\right]$$

Lower values of HL, and non-significance indicate a good fit to the data and therefore, good overall model fit. This test is performed with SPSS using Binary Logistic Regression tools too.

Conclusions

Sample size planning is important, and almost always difficult. It requires care in eliciting objectives and in obtaining suitable quantitative information prior to the study. In this paper the problems which are confronted with insufficient historical data availability is examined. To solve these problems already developed methods of determining the sample size and the

number of variables, statistical tests evaluating the accuracy of the model have been addressed. We developed graphic model determinate relationship between sample size (count of defaults), number of covariates and level of model power. Using our approach it is possible to evaluate sample size necessary to develop model with high predictive power. Having statistically good model we have qualitative internal rating system and precise measurement for the capital ratio of a bank.

References

1. Green, S.B. How many subjects does it take to do a regression analysis? Multivariate Behavioural Research, 26, 3, 499-510 – 1991
2. Harris, R.J. A primer of multivariate statistics, 2nd Edition. New York - 1985
3. Hosmer, D.W. and Lemeshow, S. Applied Logistic Regression. Second edition. John Wiley and Sons, New York - 2000
4. Nunnally, J.C. Psychometric Theory, 2nd Edition. New York: McGraw-Hill - 1978
5. Peduzzi, P., Concato, J., Kemper E., Holford, T.R. and Feinstein, A.R. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49: 1372-1379. - 1996.
6. Tabachnik, B.G. and Fidell, L.S. Using Multivariate Statistics 3rd Edition. London: Allyn and Bacon - 1989
7. Tabachnik, B.G. and Fidell, L.S. Using Multivariate Statistics 4th Edition. London: Allyn and Bacon – 2001