# 基于 Logistic 回归分析的违约概率预测研究

于立勇[1], 詹捷辉[2]
(1. 北京大学 光华管理学院, 北京 100871;
2. 哈尔滨工业大学 金融研究所, 黑龙江 哈尔滨 157001)

摘 要: 内部评级法是巴塞尔新资本协议的核心内容之一而计算客户违约概率(PD)是实施内部评级法的关键步骤。文章在结合我国国有商业银行实际数据的基础上, 利用正向逐步选择法(forward stepwise)构建了较为科学的信用风险评估指标体系, 通过 Logistic 回归模型构建了违约概率的测算模型。实证结果表明, 模型可以作为较为理想的预测工具。

关键词: 内部评级法; 违约概率; Logistic
中图分类号: F830.5 文献标识码: A 文章编号: 1001-9952(2004)09-0015-09

## 一、引 言

新巴塞尔资本协议的核心之一为内部风险评级体系, 从发达国家国际性大银行的经验看, 内部评级对于信用风险管理的作用是巨大的。新巴塞尔协议内部评级法又可以分为基础法和高级法, 而两者需要共同考虑的风险因素为违约概率(PD)。违约概率是指借款人未来一定时期内不能按合同要求偿还贷款本息或履行相关义务的可能性。在新资本协议中, "违约概率"被具体定义为借款人 1 年内的累计违约概率与 3 个基本点中的高者。巴塞尔委员会设定 0.03%的下限既是给风险权重设定下限, 同时也是考虑到银行在检验概率时所面临的困难。巴塞尔委员会在第三次征求意见稿中对客户的违约定义为: 若出现以下一种情况或同时出现以下两种情况, 债务人将被视为违约。

(1)银行认定, 除非采取追索措施, 如变现抵押品(如果存在的话), 借款人可能无法全额偿还对银行集团的债务;

(2)债务人对于银行集团的实质性信贷债务逾期 90 天以上。若客户违反

收稿日期: 2004-06-15
基金项目: 国家自然科学基金"WTO 与中国商业银行的改革与创新"(70373012)
作者简介: 于立勇(1974—), 男, 山东黄县人, 北京大学光华管理学院博士后流动站研究人员;
詹捷辉(1979—), 男, 哈尔滨工业大学金融研究所助理研究员。

了规定的透支限额或者新核定的限额小于目前的余额，各项透支将被视为逾期。

上述标准只是一个参考定义，为了选取样本和建立判别模型，还必须制定一个切实可行的违约与非违约企业的界定标准。企业违约集中和突出地表现为财务违约。以违约、无偿付能力或破产为显著特征和具体表现形式。从企业财务违约表现入手，通过分析财务违约的显著特征，就可以对企业是否违约进行准确划分。违约、无偿付能力或破产在实务中都表现为企业无法按贷款合同约定偿还银行本金和利息，因此企业能否按时偿还银行贷款本息可以作为企业违约与否的界定标准。

从统计学角度看，常用来对企业信用风险进行分析的数学工具主要包括判别分析、Logistic 回归分析、主成分分析和神经网络等四种类型。主成分分析可以从变量的相互影响关系中提取出主要因素，并根据各要素所含信息的多少确定变量关系和计算方法，一般不能单独使用，而是用来做数据的预处理；神经网络扬弃了传统预测函数的变量是线性并且相互独立的假设，能深入挖掘预测变量之间隐藏的关系，正在成为非线性违约预测函数的重要工具，但违约概率不是可以直接观察的，不能直接用来作为神经网络的学习样本；判别分析中的 Bayes 判别分析和 Logistic 回归分析均可用来进行违约概率分析，但 Bayes 判别分析需要对所研究的对象已有一定的认识，即需要用到先验概率，而国内银行信用风险度量为时不长，缺乏相应的数据积累，这种先验概率缺乏充足的说服力，如果给定的先验概率获取较为困难，Bayes 判别法可能会导致错误的结论。Logistic 回归分析是一种非线性分类的统计方法，也适用于因变量中存在定性指标的问题，而且 Logistic 判别函数的建立方法——极大似然估计法有很好的统计特性。本文尝试用 Logistic 回归模型来研究违约概率，以期为定量衡量信用风险提供一种建模方法。

## 二、Logistic 模型与信用风险评估

线性回归模型(linear regression model)在定量分析中是非常流行的统计分析方法，但在考虑计算 PD 模型时，由于因变量是一个二分类变量(“正常”或者“违约”，也可记为“0”与“1”)，而不是一个连续变量，所以对于二分类因变量的分析需要使用非线性函数。

事件发生的条件概率 $P(y_i=1|x_i)$与 $x_i$ 之间的非线性关系通常是单调函数，即随着 $x_i$ 的增加单调增加或者减少。一个自然的选择便是值域在(0,1)之间有着 S 形状的曲线，这样在 $x_i$ 趋近于负无穷时有 $E(y_i)$趋近于 0，在 $x_i$ 趋近于正无穷时有 $E(y_i)$趋近于 1。这种曲线类似于一个随机变量的累积分布曲线。在二分类因变量分析中曾使用多种分布函数，最常用的函数是 logistic 分布。

假设有一个理论上存在的连续反应变量 $y_i^*$ 代表事件发生的可能性，其值域为负无穷至正无穷。当该变量值跨越一个临界点 c(不妨令 c=0)，便导

致事件发生，于是有：

当 $y_i^* < 0$ 时，$y_i = 1$；其他，$y_i = 0$。

这里，$y_i$ 是实际观察到的反应变量。$y_i = 1$ 表示事件发生；$y_i = 0$ 表示事件未发生。如果假设 $y_i^*$ 和自变量 $x_i$ 之间存在一种线性关系，即：

$$y_i^* = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

则：

$$P(y_i = 1 | x_i) = P[(\alpha + \beta x_i + \varepsilon_i) > 0] = P[\varepsilon_i > (-\alpha - \beta x_i)] \tag{2}$$

为了取得一个累积分布函数，对上式做如下处理，由于 Logistic 分布和正态分布都是对称的，因此：

$$P(y_i = 1 | x_i) = P[\varepsilon_i \leqslant (\alpha + \beta x_i)] = F(\alpha + \beta x_i) \tag{3}$$

其中，F 为 $\varepsilon_i$ 的累积分布函数，分布函数的形式依赖于 $\varepsilon_i$ 的假设分布。标准 Logistic 分布的平均值为 0，方差等于 $\pi^2/3 \approx 3.29$，之所以选择这样一个方差是因为它可以使累积分布函数取得一个较为简单的公式：

$$P(y_i = 1 | x_i) = P[\varepsilon_i \leqslant (\alpha + \beta x_i)] = \frac{1}{1 + e^{\varepsilon_i}} \tag{4}$$

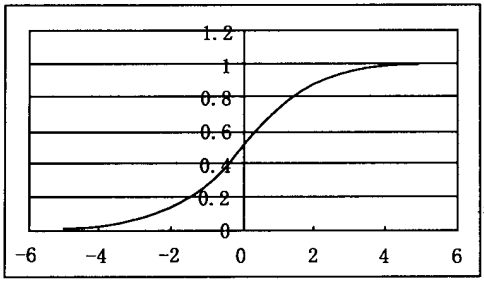


**图 1　Logistic 函数的曲线图**

这一函数称为 Logistic 函数，具有 S 型分布，见图 1。

在这一图形的左侧，当 $\varepsilon_i$ 趋近于负无穷时，Logistic 函数有：

$$P(y_i = 1 | x_i) = \frac{1}{1 + e^{-(-\infty)}} = \frac{1}{1 + e^{\infty}} = 0 \tag{5}$$

与此相对，当 $\varepsilon_i$ 趋近于正无穷时，Logistic 函数有：

$$P(y_i = 1 | x_i) = \frac{1}{1 + e^{(-\infty)}} = \frac{1}{1 + e^{-\infty}} = 1 \tag{6}$$

无论 $\varepsilon_i$ 取任何值，Logistic 函数 $P(y_i = 1 | x_i) = \frac{1}{1 + e^{-\varepsilon_i}}$ 的取值范围均在 0 至 1 之间。Logistic 函数这一性质保证了由 Logistic 模型估计的概率不会大于 1 或小于 0。Logistic 函数的另一个性质也很重要，即这个函数的形状适用于研究概率。如图 1 所示，如果从 $\varepsilon_i = -\infty$ 开始向右移动，当 $\varepsilon_i$ 增加时，这一函数的值先是很缓慢地增加，然后转向迅速增加，之后增加的速度又开始逐渐减缓，最后当 $\varepsilon_i$ 趋近于 $+\infty$ 时，函数值趋近于 1。Logistic 函数的 S 型曲线表明，$\varepsilon_i$ 的作用对于某个样本发生某一事件的可能性是变化的，在 $\varepsilon_i$ 很小时其

作用也很小，然而在中间阶段对应的可能性增加很快，但是在 $\varepsilon_i$ 值增加到一定程度后，可能性就保持在几乎不变的水平了。这说明，$\varepsilon_i$ 在 $P(y_i=1|x_i)$接近于 0 或 1 时的作用要小于当 $P(y_i=1|x_i)$处于中间阶段时的作用。这种非线性函数的形式有助于解决线性概率模型所不能解决的问题。比如，在企业违约问题中，净资产收益率对企业违约的影响，并不一定净资产收益率增加到一定量，非违约概率就会固定地增加到一定量。实际的情况是，净资产收益率在某一段水平内变化时对违约概率影响较大，而较低或较高的净资产收益率对违约概率的变化影响都不大。

由 Logistic 函数到基于 Logistic 回归分析的信用风险评估模型，首先需要重新定义 $\varepsilon_i$，此时，$\varepsilon_i$ 被定义为一系列影响违约概率因素的线性组合，即：

$$\varepsilon_i = \alpha + \sum_{k=1}^{m}\beta_k x_{ki},\ P(y_i = 1 \mid x_i) = \frac{1}{1+\exp\left[-\left(\alpha+\sum_{k=1}^{m}\beta_k x_{ki}\right)\right]} \qquad (7)$$

上述的非线性函数用 Logit 变换可以转变为线性函数：

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{k=1}^{m}\beta_k x_{ki} \qquad (8)$$

将 LogitP 看成因变量，Logistic 回归就与多元线性回归模型形式是一致的，不同的是：(1)Logistic 回归模型中因变量 y 是二分类的，而不是连续的，其误差的分布不再是正态分布而是二项分布，且所有的分析均建立在二项分布的基础上。(2)也正是基于上述原因，Logistic 回归系数的估计不再用最小二乘法，而要用极大似然法。系数及模型检验也不是 t 检验和 F 检验，而要用似然比检验和 Wald 检验等。

## 三、信用风险评估指标体系的确立

通过综合考虑信用风险的各影响因素，借鉴我国财政部统计评价司的企业效绩评价指标体系和国有商业银行企业资信评估指标体系以及国内外有关文献的相关指标，在分类、汇总、整理的基础上，同时兼顾数据的可获取性原则和可量化原则，依次选取经济性质、流动比率、速动比率、超速动比率、营运资金/总资产、资产负债率、流动资产周转率、有形净值债务率、营运资本负债率、净资产收益率、资产收益率、销售净利率、销售收入/总资产、销售毛利率、营运资金/销售净收入、存货周转率、应收账款周转率、总资产周转率、产权比率、固定资产周转率等 21 项指标。通过这些指标可以较为全面地反映企业的盈利能力、偿债能力、运营效率和盈利能力等层面的信息。同时，也应该看到这些指标之间存在一定的相关性与可替代性，需要在一定统计水平上加以挑选。

常用的选择方法有：(1)正向逐步选择法(forward stepwise)：即在截距模型的基础上，将符合所设置水平的自变量一次一个地加入模型；(2)反向逐步

选择法(backward stepwise)：在模型包括所有候选变量的基础上，将不符合保留要求显著水平的自变量一次一个地删除掉；(3)混合逐步选择法(combined stepwise)：它将正向选择和反向选择结合起来，根据所设的显著性标准分别将变量加入到模型中去或剔除掉。这种方法既可以由正向选择法开始，也可以由反向选择法开始。以上三种方法主要在设计程序上的算法不同，处理结果一般是一致的。笔者利用 SAS 完成这一过程，选用正向逐步选择法。以某国有商业银行为例，选择同一行业(制造业)的企业客户为研究对象，构建了容量为 132 个样本的样本集，其中包括 35 个正常类贷款企业和 97 个发生不同贷款损失的违约类贷款企业。

在正向逐步选择过程中，Score 统计量用来做加入选择，Wald 统计量用来做删除选择。在正向逐步选择的第 0 步(Step 0)，只有一个常数(即截距 intercept)加入模型。残差 $\chi^2$ 统计值可以用来检验所有不在模型中的变量系数都为 0 的零假设。由于残差 $\chi^2$ 的 p 值很小($p<0.0001$)，表明至少有一个为非0 值的系数。从 Analysis of Effects Not in the Model 的表中，可以看出，$X_6$ 有最大的 Score 统计值，其对应的 p 为0.0839，小于加入标准 0.2，所以进入模型。按照同样的过程，$X_1$，$X_3$，$X_{14}$，$X_4$ 进入模型，输出结果见表 1。

表 1　正向选择过程参数表

| Summary of Forward Selection | | | | | |
|---|---|---|---|---|---|
| Step | Effect Entered | DF | Number In | Score Chi－Square | Pr＞ChiSq |
| 1 | $x_6$ | 1 | 1 | 2.9869 | 0.0839 |
| 2 | $x_1$ | 1 | 2 | 2.9896 | 0.0838 |
| 3 | $x_3$ | 1 | 3 | 2.9087 | 0.0881 |
| 4 | $x_{14}$ | 1 | 4 | 2.7919 | 0.0947 |
| 5 | $x_4$ | 1 | 5 | 3.0247 | 0.082 |

需要做出说明的是，以“select＝0.2”为筛选变量的标准，没有采用常用的检验标准，比如 0.05，原因在于如果不选择大一点的显著水平，就有可能遗漏掉某些重要的自变量。他们很有可能在简单分析时显示与结果变量的弱相关，而在多元分析时就成为重要的自变量。所以选择一个足够大的水平，以保证将有可能成为重要预测变量的候选者都纳入到多元分析中。

由于变量 $X_7$ 的 Pr＞ChiSqure 值为0.2139，与筛选标准 0.2 相差不大，同时考虑到解释变量的充分性，也将其纳入到模型中。

经过以上测算与分析，筛选出 6 个自变量，分别为：经济性质、速动比率、超速动比率、资产负债率、流动资产周转率和净资产收益率。剔除掉与速动比率相关性较强的超速动比率后，剩余 5 个自变量，这些指标涉及了经济性质、运营效率、偿债能力、盈利能力等四个方面的内容，可以较为科学地反映贷款企业的信贷风险。

## 四、违约概率的测算

综合考虑了 Logistic 回归模型对样本构成的要求，在上文构造的训练样本集基础上，构建了容量为 51 个贷款企业的测试样本集，其中包括 11 个正常类贷款企业和 40 个发生不同贷款损失的违约类贷款企业。

在处理上，首先直接把 5 个自变量纳入模型，经级大似然估计得出的系数存在较大的标准误差，且某些指标不能通过统计检验。通过分析 5 个自变量，发现偿债能力指标与违约风险之间并非存在完全的线性关系，即资产负债率对企业财务状况的影响，并非资产负债率越高越好，也不是越低越好。实际情况是，合适的资产负债率有助于财务的稳健性并起到有利的财务杠杆作用。过高意味着高举债经营，财务风险偏高，不够稳健；过低则未能充分发挥财务杠杆的作用，未能达到企业价值最大化，因此资产负债率对企业违约风险的影响应该是非线性的，考虑纳入其平方值，使其对因变量的影响变为开口朝下的抛物线形状。

表 2　**Pearson** 卡方和 **Deviance** 拟合优度检验

| Deviance and Pearson Goodness－of－Fit Statistics | | | | |
|---|---|---|---|---|
| Criterion | DF | Value | Value/ DF | Pr> Chi－Square |
| Deviance | 125 | 136. 3 | 1. 090 2 | 0. 231 2 |
| Pearson | 125 | 131. 2 | 1. 049 9 | 0. 333 5 |
| Number of unique profiles: 132 | | | | |

由表 2 可知，Pearson 卡方和 Deviance 统计量结果都表明统计显著性不是很强。在本文的算例中，由于自变量含有连续变量，协变类型数量很大，因此每个协变类型所对应的观察案例并不多，指标 Deviance 和 Pearson 卡方不能有效评估拟合优度，所以采用 Hosmer－Lemeshow（HL）检验（见表 3）。该方法根据模型预测概率的大小将数据分成规模大致相同的 10 个组，然后根据每一组中因变量各种取值的实测值与理论值计算 Pearson 卡方。通常用于自变量很多，或者自变量中包含连续性变量的情况。HL 的检验结果见表 3，p＝0. 5772，统计不显著，不能拒绝关于模型拟合数据较好的假设。

表 3　**Hosmer－Lemeshow** 拟合优度检验

| Hosmer and Lemeshow Goodness－of－Fit Test | | | | | |
|---|---|---|---|---|---|
| | | Y＝1 | | Y＝0 | |
| Group | Total | Observed | Expected | Observed | Expected |
| 1 | 13 | 4 | 5. 48 | 9 | 7. 52 |
| 2 | 13 | 7 | 7. 37 | 6 | 5. 63 |
| 3 | 13 | 9 | 8. 37 | 4 | 4. 63 |
| 4 | 13 | 11 | 9. 11 | 2 | 3. 89 |
| 5 | 13 | 10 | 9. 50 | 3 | 3. 50 |
| 6 | 13 | 11 | 9. 94 | 2 | 3. 06 |
| 7 | 13 | 10 | 10. 39 | 3 | 2. 61 |
| 8 | 13 | 10 | 11. 02 | 3 | 1. 98 |
| 9 | 13 | 10 | 11. 66 | 3 | 1. 34 |
| 10 | 15 | 15 | 14. 17 | 0 | 0. 83 |
| Goodness－of－fit Statistic＝6. 6287 with 8DF（p＝0. 577 2） | | | | | |

模型 $\chi^2$ 统计(Model Chi—Square Statistic),定义为零假设模型与所设模型之间在—2LL̂ 上的差距。LL̂ 为模型的最大似然值取对数,似然比统计量近似地服从 $\chi^2$ 分布(Hanushek 和 Jackson, 1977; Aldrich 和 Nelson, 1984; Greene, 1990)。似然比统计量如下:

$$Gs=-2\ln\left(\frac{\hat{L}_0}{\hat{L}_s}\right)=-2(\ln\hat{L}_0-\ln\hat{L}_s)=2\hat{LL}_s-2\hat{LL}_0 \qquad (9)$$

实际上,模型 $\chi^2$ 检验与多元线性回归中的 F 检验十分类似,这里零假设为除常数项外的所有系数都等于 0。从表 4 可以看出,显著性水平为0.0117,模型 $\chi^2$ 统计较为显著,所以认为自变量所提供的信息是有用的。

表 4　模型卡方统计以及信息测量指标

| Model Fitting Information and Testing Global Null Hypothesis BETA=0 | | | |
|---|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates | Chi—Square for Covariates |
| AIC | 154.691 | 150.281 | |
| SC | 157.574 | 170.460 | |
| —2LOGL | 152.691 | 136.281 | 16.441 with 6DF (p=0.0117) |

类似于线性回归中的确定系数,R—Square 为一般线性模型的确定系数,Max—rescaled R—Square 为回归的调整类确定系数,模型输出中 R—Square 为0.6301,Max—rescaled R—Square 为0.8134。数据变异中被解释的比例为81.34%。

在线性回归中,估计未知总体参数时主要采用最小二乘法,极大似然估计法是统计分析中另一常用模型参数估计方法。与最小二乘法相比,极大似然估计法既可以用于线性模型,也可以用于更为复杂的非线性估计。由于 Logistic 回归是非线性模型,因此本文采用极大似然估计方法,结果见表 5。

表 5　极大似然估计分析

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | Wald Chi—Square | Pr> Chi—Square |
| INTERCPT | 1 | 5.447 | 2.8463 | 3.6623 | 0.0557 |
| $X_1$ | 1 | 0.998 | 0.5142 | 3.7668 | 0.0523 |
| $X_2$ | 1 | —1.3165 | 0.7248 | 3.2995 | 0.0693 |
| $X_3$ | 1 | —11.4773 | 5.6434 | 4.1361 | 0.042 |
| $X_4$ | 1 | 4.9061 | 2.8085 | 3.0516 | 0.0807 |
| $X_5$ | 1 | 0.0263 | 0.0178 | 2.1832 | 0.1395 |
| $X_6$ | 1 | —2.003 | 1.1173 | 3.2139 | 0.073 |

对 Logistic 回归模型与预测准确性之间的第二种测量方法是建立在观测的反应变量和模型预测的条件事件概率之间的关联基础上。序次相关指标(rank correlation index)测算结果见表 6,共有3 395对数据对,其中和谐的占70.7%,不和谐的占29.0%,其他占0.3%。

表 6　预测概率与观测值之间的关联

| Association of Predicted Probabilities and Observed Responses | | | |
|---|---|---|---|
| Concordant | 70.7% | Somers' D | 0.417 |
| Discordant | 29.0% | Gamma | 0.418 |
| Tied | 0.3% | Tau－a | 0.164 |
| (3395 pairs) | | c | 0.708 |

## 五、测试结果及分析

基于上述分析与测算，可得出 Logistic 回归分析方程为：

$$LogitP = 5.447 + 0.998x_1 - 1.317x_2 - 11.477x_3 + 4.906x_4 + 0.026x_5 - 2.003x_6 \quad (10)$$

自变量分别为：$x_1$ 经济性质，$x_2$ 速动比率，$x_3$ 资产负债率，$x_4$ 资产负债率的平方，$x_5$ 流动资产周转率，$x_6$ 净资产收益率。通过 Logit 变换，把测试样本带入方程，即可得出属于正常组及违约组的概率，样本预测准确性见表 7。正常组正确判别率为 80%，违约组正确判别率为 92.7%。由于篇幅有限，各个企业具体发生违约概率未能给出。

表 7　预测结果

| Y | Predict | | |
|---|---|---|---|
| Frequency | 0 | 1 | Total |
| 0 | 8 | 2 | 10 |
| | 80% | 20% | 100% |
| 1 | 3 | 38 | 41 |
| | 7.3% | 92.7% | 100% |
| Total | 11 | 40 | 51 |

本文系统探讨了 Logistic 函数与模型作为预测贷款企业违约概率的理论基础，并给出了构建模型解释变量的统计方法及分析方法，同时结合商业银行实际数据运用 Logistic 回归模型对企业违约概率进行了实证分析。研究结果表明，Logistic 模型是一种较为理想的企业违约概率预测工具。然而，模型的实证过程中也存在一定的不足，如出于对指标量化的考虑，本文所构建的指标体系主要集中于财务指标，尚未充分考虑非财务因素对贷款企业信用风险的影响与作用。Logistic 模型自身也存在一定不足，如对线性可分的样本不可采用级大似然估计，样本的数量不宜太少，这些都是需要进一步研究改进的方向。

参考文献：

[1] Cebenoyan, A. Sinan; Strahan, Philip E. Risk management, capital structure and lending at banks[J]. Journal of Banking and Finance. 2004, 28(1): 19~43.

[2] Murphy, Austin. An empirical analysis of the structure of credit risk premiums in the Eurobond market[J]. Journal of International Money and Finance. 2003, 22(6): 865~885.
[3] 于立勇, 曹凤岐. 论新巴塞尔资本协议与我国银行资本充足水平[J]. 数量经济技术经济研究. 2004, (1).
[4] 于立勇. 商业银行信用风险评估预测模型的研究[J]. 管理科学学报. 2003, (10).
[5] 于立勇. 基于具有吸收态马尔可夫链的商业银行信贷风险管理研究[J]. 数量经济技术经济研究. 2000, (1).
[6] 王春峰, 万海晖, 张维. 组合预测在商业银行信用风险评估中的应用[J]. 管理工程学报, 1999, (1).
[7] 于立勇. 信用风险评估中贷款风险度的波动性分析[J]. 数量经济技术经济研究, 2002, (3).
[8] 于立勇. 商业银行信用风险衡量的一种新标准[J]. 数量经济技术经济研究. 2002, (9).
[9] 沈沛龙, 任若恩. 新巴塞尔协议资本充足率计算方法剖析[J]. 金融研究, 2002, (6).
[10] 张玲, 曾维火. 基于 Z 值模型的我国上市公司信用评级研究[J]. 财经研究, 2004, (6).
[11] 张维, 李玉霜. 商业银行信用风险分析综述[J]. 管理科学学报, 1998, (9).
[12] 王春峰, 康莉. 基于遗传规划方法的商业银行信用风险评估模型[J]. 系统工程理论与实践, 2001, (2).

# A Research on Probability of Default prediction Based on Logistic Regression Analysis

YU Li-yong[1], ZHAN Jie-hui[2]

(1. *Guanghua School of Management, Peking University, Beijing* 100871, *China*;
2. *Harbin Institute of Technology, Harbin* 150001, *China*)

**Abstract**: Internal rating-based approach is one of the main contents of New Basel Accord, while calculating clients' probability of default is a key procedure of practicing internal rating-based approach. Based on the practical data of China's state-owned commercial banks, this paper constructs a rather scientific credit risk evaluating system by forward stepwise, and predicting models of probability of default by logistic regression model. Experimental results prove that this model can serve as an ideal predicting instrument.

**Key words**: internal rating based approach; probability of default; logistic