基于因子分析的logistic违约概率测算模型研究[1]

彭建刚 ，屠海波 ，何婧
湖南大学 金融学院，金融管理研究中心，湖南 长沙（410079）
E-mail：pengjiangang@hotmail.com

摘 要： 本文针对一般 Logistic 违约率模型中原始数据信息的丢失、多重共线性以及没有考虑时间因素等问题，提出了基于因子分析的 logistic 违约概率测算模型。通过引入因子分析和对指标作时间加权化处理等方法改进了一般 logistic 违约概率测算模型，然后利用中国上市公司数据展开实证研究。基于因子分析的 logistic 违约概率测算模型不仅考虑了时间因素，能够解决数据丢失和多重共线性，克服了 Cramer 问题，而且测算的准确度也较高。

关键词：违约概率，因子分析，Logistic 模型
中图分类号：F832.21 文献标识码:A

1. 引 言

美国金融市场近期遭受重创，次级抵押贷款危机已蔓延至全球金融市场。次贷危机起因于对信用风险没有引起足够的重视，在商业银行信用风险管理中，违约概率的测算居于重要地位。违约概率是指借款人在未来一定时期内不能按合同要求偿还银行贷款本息或履行相关义务的可能性(概率)，即信用风险的概率测算。对借款人进行违约概率的测算，已经被列为巴塞尔新资本协议内部评级法的关键内容，是现代商业银行信用风险管理的重要环节。巴塞尔新资本协议要求[1]，采用内部评级法的银行必须对处于风险暴露中的每一借款人进行评级，并估计其违约概率。研究现代商业银行的信用风险管理，不能不关注违约概率测算问题。

20 世纪八十年代以来，logistic回归分析法逐步取代了传统的判别分析法。作为量化企业信用风险的一种主流方法，logistic回归方法不仅灵活简便，而且它的许多前提假设比较符合经济现实和金融数据的分布规律，譬如它不要求模型变量间具有线性相关关系，不要求变量服从协方差矩阵相等和残差服从正态分布等，这使得模型的分析结果比较客观。大量实证研究表明, Logistic模型估计结果与实际数据的拟合度较高, 适用性较强[2]。于立勇（2008）[3]等在结合我国国有商业银行实际数据的基础上通过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违约概率预测工具。

最近对logistic回归方法改进的研究主要有Laitinen（2000）[4]探索了泰勒级数展开在logistic回归方法预测企业违约分析中的应用。石晓军（2007）[5]则针对一般logistic回归方法存在的难以通过Hosmer-Lememshow拟合优度检验的Cramer问题，提出了边界logistic方法。

由于用来解释违约概率的信用变量具有高相关性和高维性等特点，使得在运用 logistic 回归分析进行企业违约风险预测研究时会影响 logistic 分析的过程和结果，导致大部分原始数据信息的丢失以及估计方程中出现共线性的函数关系。而且我国正在处于经济转型时期，经济发展不够稳定。如果忽视时间因素对违约概率的影响，那么就会造成在经济景气的时期，商业银行会低估企业违约的概率，从而使得银行面临巨大的信用风险；而在经济萧条阶段又会高估企业违约的概率，从而使得银行可能失去优质客户。本文正是针对这些问题提出了基于因子分析的 logistic 违约概率测算模型，最后用 ROC 分析检验了不同模型测算违约概率的

[1]本文得到国家自然科学基金项目(编号: 70673021)的资助。

精度。

2. 基于因子分析的 Logistic 模型的基本框架

2.1 考虑了时间因素的 Logistic 模型的基本原理

首先利用Logistic模型进行违约概率测算研究的有Ohlson（1980）[6]、Zavgren（1985）[7]等。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 模型的建立方法---极大似然估计法有很好的统计特性。

在 Logistic 模型中，违约概率的测算被看作一个虚拟变量问题。所谓虚拟变量指的是一种取值为 0 或 1 的变量。在经济模型中，一些变量比如季节、民族、某项政策等都可能成为影响某个因变量的重要因素。这些变量所反映的并不是数量，而是某种性质或属性。为了研究方便，我们人为构造出一种特殊变量，即虚拟变量来把这些变量定量化，规定当该变量值取 1 时，表示存在某种性质或属性，取 0 时则表示不存在。

Logistic 模型假设因变量发生的概率与其各影响因素间呈现如下的非线性关系,

$$\Pi(X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_{\ 2}+....\beta_nX_{\ n})}} \tag{1}$$

其中 $X=(X_1,X_2,\cdots X_n)^T$ 表示解释变量，$\beta=(\beta_0,\beta_1,\beta_2,\cdots\beta_n)^T$ 是对于违约发生与否的解释变量的系数，β_0 是指常数项，$\Pi(X)$=1 表示企业违约，$\Pi(X)$=0 表示企业不违约。

由于企业的各种指标会随着时间变化而变化，如果仅仅考虑最近一年的指标，那么可能由于企业的经济周期或者偶然原因造成财务指标失真，最终使得违约概率测算的不准确。为了解决这一问题，本文提出了基于时间加权的 logistic 违约概率测算模型。

为了综合考虑 t 年财务指标，我们用 X_i 表示该周期的综合指标，X_{it} 表示指标 i 第 t 年的数值，那么令

$$X_i=\frac{TX_{iT}+(T-1)X_{iT-1}+(T-2)X_{iT-2}+.....+1X_{\mathrm{i1}}}{(T+1)T/2}=\frac{\sum_{t=1}^{T}tX_{it}}{(T+1)T/2} \tag{2}$$

再把 X_i 代入（1）中就可以得到基于时间加权的 logistic 违约概率测算模型：

$$\Pi(X)=\frac{1}{1+e^{-\beta_0-\sum_{i=1}^{n}(\beta_1\sum_{t=1}^{T}(tX_{it}/((t+1)t/2)))}} \tag{3}$$

Logistic 与一般多元线性回归模型不同之处在于: (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。在二项 Logistic 模型，似然函数等于

$$l(\beta)=\prod_{j=1}^{n}\pi(X_{\ j})^{z_j}[1-\pi(X_{\ j})]^{1-z_j}\qquad j=1,2.....n \tag{4}$$

为了求解能够使$l(\beta)$达到最大化的β，需要对$l(\beta)$分别求β，β_0的微分，得到 n+1

个似然方程式，并令其等于 0。

由于 logistic 回归分析中变量间的关系是非线性的，因此一般使用迭代算法来估计解释变量的系数 β 和常数项 β_0。

2.2 一般 logistic 回归的缺陷分析

由于用来解释违约概率的信用变量具有高相关性和高维性等特点，使得在运用 logistic 回归分析进行企业违约风险预测研究时会影响 logistic 分析的过程和结果，导致大部分原始数据信息的丢失以及估计方程中出现共线性的函数关系。具体来说，logistic 回归分析要求模型解释变量之间不能具有线性的函数关系，否则共线性的问题就会导致方程中变量系数标准差的增大。从而使得模型估计系数可靠性大幅度下降，最终利用模型测算违约概率的准确性不理想。

另一方面，在模型包括众多解释变量的情况下，logistic 回归分析的目标之一是得到预测违约概率的“节约模型”方程，这个方程需要满足（1）包括尽可能少的解释变量；（2）具有最优的度量结果（3）尽可能多地考虑原始数据的信息；（4）具有经济学意义上的说服力等条件。常用的选择方法有正向逐步选择法、反向逐步选择法、混合逐步选择法。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。这类方法的缺点主要在于其完全依赖统计方法，缺乏经济学基础；此外，还导致了大部分解释变量被剔除掉了，这使得估计方程是不完整的。

为了解决 logistic 回归所存在的共线性和原始数据丢失等问题，本文在先采用时间加权方法的基础上，再用因子分析的方法对数据行进分析，最后运用 logistic 回归分析的构建模型。

2.3 因子分析基本原理

在许多研究中，为了全面系统分析问题，都尽可能完整地搜集信息，对每个研究对象往往需测量很多变量（或称指标），人们自然希望用较少的新变量代替原来较多的旧变量，而这些新变量尽可能反映旧变量的信息。因子分析正是满足这一要求的处理多变量的方法。由于它们能浓缩信息，使指标降维，简化指标结构，使分析问题简单、直观、有效，故被广泛地应用于医学、心理学、经济学等领域。

为了尽可能精确的测算违约概率，人们一般会尽量地收集贷款的信息。如一般对公贷款除了企业自身 3 年的财务报表，还需要企业管理层、行业、地区等大量的信息，转化为指标的话一般有上百个之多。而这些指标很多是高度相关的，如果直接使用这些指标的话，不仅增大了建模的难度，也可能受一些无关的指标干扰。另外，各个指标之间的数量级差别很大，容易造成数量级较小的重要变量被低估甚至忽略。而因子分析则能在解决这些问题的同时，尽可能多的保留原始变量的信息。

因子分析的步骤包括：因子模型的构建、因子负载矩阵求解、因子旋转和因子得分的求解。因子分析的一般模型：设 x 为 $p\times1$ 随机向量，其均值为 μ，协差阵为 $\sum=\left\{\sigma_{ij}\right\}$，我们称 x 为有 k 个因子的模型，若 x 能表为：

$$x=\mu+\Lambda f+u \tag{5}$$

式中 Λ：$p\times k$ 是未知常数阵，f：$k\times1$ 和 u：$p\times1$ 为随机向量。f 称为公共因子，u 叫做特殊因子，Λ 叫因子负载矩阵。

因子负载矩阵一般可由主因子法求解得到。当我们一旦获得了公共因子和因子负载以后，我们应该反过来考察每一个样本，可以通过巴特莱特估计、贝叶斯估计估计等方法得到因子得分。

在进行违约概率测算的过程中，本文采用巴特莱特统计估计的方法，从众多反映风险财务指标中计算出包含充分指标信息的公共因子，这些公共因子比原始财务指标具有更优的统计特征，运用原始变量的组合值即因子得分作为反映信用风险的变量作进一步研究。

2.4 基于因子分析的 logistic 回归模型的优点

把由因子分析得到的向量 $f=(Z_1,Z_2,.....Z_n)$ 作为 logistic 模型的新的解释变量代替，即可以得到新的测算违约概率的模型。这个模型与一般的 logistic 模型相比在保留 logistic 模型原有优势的同时，主要有以下几个优点：

（1）.模型通过对数据标准化的处理，消除了变量间在数量级上或量纲的不同而产生的影响，每个变量的均值都为 0，方差为 1。

（2）因子的指标之间由于互不相关，这样在 logistic 回归分析中，避免出现常见的多重共线性，大大增加了 logistic 回归分析中系数的可靠性。

（3）在保留尽可能多的信息的前提下，使得 logistic 回归分析中的变量大大减少，从而在不影响违约概率测算精度的情况下显得“节约”。

（4）相对于 logistic 回归分析完全依赖统计方法的变量选择，因子分析可以更好的考虑变量的经济学意义，从而使得模型更有实用价值。

本文用基于因子分析的 logistic 回归分析对我国上市公司的财务及资本市场数据建立违约概率测算模型，并用 ROC（受验者工作特征线）的检验理论来检验模型的表现能力。

3. 我国上市公司的实证分析

3.1 数据的选取和说明

模型样本包括在深沪上市公司（包括 A 股和 B 股）共计 1629 家，考虑到了行业的特性，剔除了金融、保险公司 22 家，样本包括非 ST 公司 1446 家和 161 家 ST 公司，收集了样本公司 2004-2007 的财务数据和资本市场数据（均来自国泰君安数据库）。在去掉相关性明显很强的指标和共线性指标（即某个指标可以由其他指标线性表出）后。本文考虑了获利能力、流动性、现金流量、资产负债、资本市场等五大类 22 个指标。这 22 个指标在已有的研究中证明对违约概率的研究是有用的。本文对违约企业的定义，采用传统的分析方法，即视 ST 股(上市公司因财务状况异常而被“特殊处理”)为违约的借款企业，非 ST 股为不违约借款企业。

在已有的研究中，获利能力比是首要的指标。本文使用的获利能力比例包括总资产净利润率、营业毛利率、营业净利润率、资产报酬率、投入资本回报率。总资产净利润率（ROA）是指净利润对总资产的比，它给投资者描述了一个公司的投资资金如何有效地转换成净利润的概念，ROA 值越高，公司的资质就越好。在通常的研究中，ROA 是一个度量公司获利能力的重要指标。

财务杠杆比率也是预测公司信用风险的重要变量。本文考虑了财务杠杆系数、经营杠杆系数、综合杠杆。资产负债指标包括资产负债率、所有者权益比率、流动负债比率、长期负债比率。流动性指标包括流动比率、速动比率、营运资金比率、营运资金对资产总额比率。

现金流量指标现金流量对流动负债比率、每股经营活动现金净流量、每股筹资活动现金净流量、每股现金净流量、销售现金比率。

资本市场指标 P/S 表示的是股价和每股销售额的比率，它很多时候被认为是衡量一个公司价值的重要指标，一个具有较低 P/S 指标的公司，一般认为比较具有投资价值，反应了资本市场对公司价值看法，一定程度上能反映公司的信用风险状况。在一个经济周期内，市盈率的波动会很大(如一般来说，钢铁股当市盈率很低的时候，往往是其盈利能力下降的开始)，而市净率无法体现不同资产质量之间的差别。所以这里只选择了股价和每股销售额的比率。

3.2 因子分析过程

3.2.1 数据处理和分析

用式（2）的时间加权数据处理方法，对 04-06 的公司样本数据作时间加权平均处理为 06 的综合指标数据，对 05-07 的公司样本数据作时间加权平均处理为 07 的综合指标数据，合格的样本总数共计 3114 个（对于上市不到 3 年的公司，相应的 T 取 1，或者 2）。

首先对这些变量做两两间的皮尔逊相关性分析，我们发现这些变量之间存在显著的高度相关（Pearson 相关系数大于 0.8）及强相关（Pearson 相关系数在 0.5 到 0.8 之间），这说明对原始数据进行因子分析是很有必要的。

对数据的进行描述分析，可以发现各个变量的最小值、最大值、均值与方差有很大的差异，这种差异主要是由于各个变量间在数量级上或量纲上的不同，这会对后续分析产生不利的影响。为了消除这种影响，我们通过把所有变量都变为均值为 0、方差为 1 的方法（即用原始数值减去均值，再除以方差），先对原始数据作了标准化处理。

3.2.2 变量共同度分析

对变量进行共同度分析可知，除了销售现金比率和 P/S 比率,其它变量的共同度对前几个因子（特征值大于 1）均在 0.8 以上，这表明大部分变量都很好被前几个因子所解释。

3.2.3 特征值分析和因子矩阵

对数据进行特征值分析，我们发现变量相关阵前 10 个因子的特征根均大于 1，它们一起解释了总方差的 93.761%（累积贡献率）。这说明这 10 个因子提供了原始数据的足够信息。从碎石图也可以看出，前 10 个主成分的特征值大于 1，且明显大于后面主成分的特征值。这说明因子分析结果是比较理想的。

由初始因子负荷矩阵得到的旋转以后的因子矩阵可以很清晰的得出各个主成分与原始变量的关系：

表 1　旋转因子矩阵
Tab.1 Rotating factor matrix

	Component									
	1	2	3	4	5	6	7	8	9	10
流动比率	.018	.017	.022	.008	3.05E-005	.990	.011	.057	-.029	-.036
速动比率	.013	.010	.019	.006	.004	.991	.008	.051	-.020	.009
营运资金比率	.967	.109	-.011	.022	-.004	.032	.001	.012	.003	.004
营运资金对资产总额比率	.999	-.008	.005	.001	-.004	.009	.000	.004	.000	.001
资产负债率	-.005	-.003	.002	-1.000	.000	-.007	.000	-.006	.000	-.002

所有者权益比率	.005	.003	-.002	1.000	.000	.007	.000	.006	.000	.002
流动负债比率	-.004	.002	-.999	.002	-.005	-.020	.000	-.022	-.015	-.012
长期负债比率	.004	-.002	.999	-.002	.005	.020	.000	.022	.015	.012
营业毛利率	.022	.979	-.002	.001	-.003	.010	.000	.004	.011	-.001
营业收入净利润率	.031	.951	-.003	.000	-.007	.007	.000	.003	.017	-.004
资产报酬率 A	.995	-.011	.009	-.006	.004	-.002	.000	-.001	.001	.002
总资产净利润率	.995	-.014	.009	-.006	-.004	-.002	5.10E-005	-.001	.001	.002
投入资本回报率	.014	.006	.012	.001	-.981	.025	-.001	-.002	-.046	.044
财务杠杆系数	.003	-.014	.031	.000	-.109	-.041	.029	-.022	.896	-.084
经营杠杆系数	.001	.001	.005	-.001	.001	.023	.948	-.013	-.011	.000
综合杠杆	.000	.000	-.004	.000	-.002	-.005	.949	.003	.022	-.003
P/S	-.011	-.849	-.002	-.004	-.010	-.010	.000	-.005	.031	-.018
现金流量对流动负债比率	.007	.006	.023	.002	.977	.030	-.001	-.002	.044	.096
每股经营活动现金净流量	.007	.014	.030	.004	.045	-.028	-.003	.000	.056	.971
每股筹资活动现金净流量	.007	.010	.083	.007	-.003	.026	-.008	.874	-.053	-.379
每股现金净流量	.006	.004	-.027	.006	.003	.104	-.005	.888	-.006	.348
销售现金比率	.002	.008	-.005	.000	.225	-.005	-.020	-.026	.842	.163

第一主成分主要由营运资金比率、营运资金对资产总额比率、资产报酬率、总资产净利润率，这反映了公司的盈利能力和流动性能力，这两种能力最能放映公司信用风险状况也符合实际情况，这一指标也具有很强的经济学意义。

第二主成分主要由营业毛利率、营业收入净利润率、P/S 组成，它反应了公司销售、运营效率和定价能力或策略，这与第一主成分相比，各有侧重点。

第三主成分主要由流动负债比率、长期负债比率构成，反应了公司的偿债能力。

第四主成分主要资产负债率、所有者权益比率构成。主要反应了公司的融资结构。

第五主成分由投入资本回报率、现金流量对流动负债比率构成。

第六主成分分别主要由流动比率和速动比率得出，反应了公司的短期流动性。

第七主成分主要由经营杠杆系数，综合杠杆构成，反应了公司的杠杆经营程度。

第八、第九第十主成分则主要由现金流指标组成，反应了公司现金流入流出情况。

从上面的结果我们可以很清楚地看到盈利能力、偿债能力、融资结构这些是关系到公司是否违约的最重要的指标，而其它如杠杆经营程度、现金流量等指标也起着一定作用。

如果简单把指标分成几类，那么无法避免各个类别之间存在相关性，而且指标分类的人为性较大，而因子分析就可以在不失去经济学意义的前提，更加科学的处理指标。

根据因子负荷矩阵和各个矩阵的特征值，可以得到每个公司的各个因子值。我们以第一主因子为例，说明求解过程。对因子负荷矩阵第一列的因子负荷值分别除特征值的平方根得到新的因子负荷值，以这个因子为权重对某个公司的各个指标加权求和，即可以得到该公司第一主成分的值。

3.3 logistic 回归结果

计算出因子值，再把这些数值作为 logistic 回归的解释变量。可以通过 SPSS13.0 软件得到以下结果：

3.3.1 模型的统计检验

为了评估拟合优度,我们采用 Hosmer - Lemeshow (HL) 检验。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 的检验的卡方统计量为 62.310,统计显著,接受关于模型拟合数据较好的假设。这说明基于因子分析的 logistic 回归方法避免了一般 logistic 回归无法通过 Hosmer - Lemeshow 检验的 Cramer 问题。

模型 χ^2 统计,定义为零假设模型与所设模型之最大对数似然值之差,似然比统计量近似地服从 χ^2 。实际上,模型 χ^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 。检验统计量均为 406.331（显著水平小于 0.1%）可以看到,模型的估计是显著的。

模型的拟合优度的检验统计量 Cox & Snell R Square、 Nagelkerke R Square 检验分别达到 0.558、0.745，可见模型的拟合优度还是比较理想的。

3.3.2 模型系数估计

模型系数的极大似然估计结果见下表：

表 2 方程系数估计结果变量
Tab.2 Variable of the equation coefficient estimates

	B	S.E.	Wald	df	Sig.	Exp(B)
FAC1_2	-12.003	2.683	20.013	1	.000	.000
FAC2_2	-18.380	3.451	28.371	1	.000	.000
FAC6_2	-1.250	.250	24.946	1	.000	.286
FAC8_2	-.912	.115	62.533	1	.000	.402
FAC10_2	-.539	.085	40.270	1	.000	.583
Constant	-1.743	.156	125.492	1	.000	.175

由表2可以得到最终的logistic模型为：

$$\log(P/1-P) = -12.003 * fac1 - 18.380 * fac2 - 1.250 * fac6 - 0.912 * fac8 - 0.539\, fac10 - 1.743$$

从模型结果可以看到，第一因子和第二因子的系数明显比其它因子的系数要大，且 Wald 统计量分别为 20.013、28.371，说明变量的作用是显著。这与因子分析的结果是一致的。从模型的系数估计,我们可以发现目前影响我国上市公司违约概率的主要是盈利能力、流动性能力、公司的运营效率和定价能力。

3.3.3 违约概率的散点图

为了考察违约概率的分布情况，我们可以看到所有样本违约概率的散点图，可以看到违约概率集中于平均违约率 11%附近，且图形呈现偏峰厚尾的现象。这与现行的违约概率分布的研究结果是一致的。

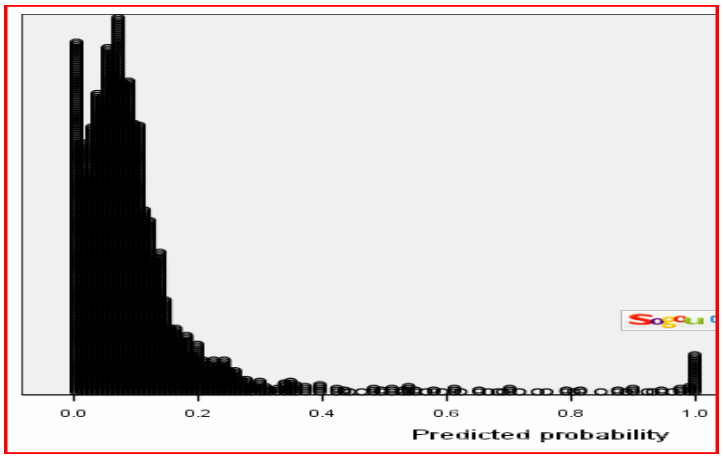

图 1 违约概率散点图

Fig1 Default probability plot

3.4 与一般 Logistic 模型的 ROC 比较分析

现有的多元判别分析、支持向量机分类（SVM）、logistic 分类等方法均采用给定置信水平，再通过比较最终结果犯这两类错误的多少来验证其有效性。这类检验方法是一种静态的方法，它假设贷款的违约概率超过一定临界值即视为违约，其结果依赖于临界值选择，而实际上贷款的违约概率超过一定临界值并不意味着违约，也就是说这样的临界值在实际业务中是不存在的。那么这类方法有效性就很值得怀疑了。针对这一问题，本文引入的 ROC 分析方法能有效的解决这一问题。ROC 分析本质上是一种动态的反应 α 型错误和 β 型错误的检验方法。

受验者工作特征（Receiver Operating characteristic）最初起源于更好地理解无线电接受器的信号噪音率。自Lusted[8]等首先将ROC分析应用于医学诊断中后，其价值日益受到广泛重视，在经过大量的研究及临床实践，现已成为临床科研文献中应用较为广泛的统计方法[9]。ROC分析的本质就是动态分析、比较不同试验在多个诊断阈值条件下，其相对应的敏感性及特异性曲线的差异，并以AUC（Area Under Curve）值作为评价ROC曲线特性的参数，AUC值在 0．5～1．0 之间有价值，A 值越接近 1．0，其价值越高。

下面我们对基于因子分析的 logistic 违约概率测算模型（以下简称方法 1）和一般 logistic 模型作 ROC 比较分析。首先计算了全部指标的一般 logistic 模型，由于原始变量之间相关性和高维性，使得在估计 logistic 模型时精度不够，无法正常的估计出违约概率。为了解决这一问题，我们用去除相关性强的指标的一般 logistic 模型（由于在去除相关性时存在人为性，为了得到一般性结果，我们采用了选择了两种不同的人为剔除强相关性指标后的结果，分别使用一般的 logistic 模型,以下简称方法 2、3）作为比较的对象。对这三种方法在不同选择方法下测得的违约概率分别做 ROC 分析可以得到：

表 3　　曲线下面积

Tab.3　Area Under the Curve

		面积	标准误	显著水平
全部选择	方法1	0.836	0.13	0.00

	方法2	0.826	0.14	0.00
	方法3	0.816	0.15	0.00
前向选择法	方法1	0.844	0.13	0.00
	方法2	0.807	0.14	0.00
	方法3	0.817	0.15	0.00
后向选择法	方法1	0.844	0.13	0.00
	方法2	0.807	0.14	0.00
	方法3	0.817	0.15	0.00

从不同的 AUC 值以可以看出，基于因子分析的 logistic 回归模型 AUC 值都比较高，因此我们可以得出结论，基于因子分析的 logistic 模型在违约概率测算的精度比一般的 logistic 模型相对较高。

4. 结论

1. 一般的 logistic 回归方法由于用来解释违约概率的信用变量具有高相关性和高维性等特点，使得利用模型测算违约概率的准确性不理想。与一般 logistic 回归方法相比，本文提出的基于因子分析的 logistic 违约概率测算模型具有以下优点：加入了对时间加权的方法,考虑了时间周期的影响；与一般的解决相关性的直接去除变量的方法相比，本文的方法没有主观性，不会因人而异，而且计算程序化、标准化，易于实际操作；在数据复杂繁多的情况下，本文的方法也可以在不丢失变量的同时使得模型显得节约，扩大了 logistic 模型测算违约概率的应用范围。

2. 从实证结果来看，基于因子分析的 logistic 违约概率测算模型不仅能解决数据丢失和多重共线性，拟合效果较好，克服了一般 logistic 模型不能通过 Hosmer-Lememshow 拟合优度检验的 Cramer 问题，而且通过与一般 logistic 模型的 ROC 方法比较，发现基于因子分析的 logistic 模型测算违约概率的准确度更高。

3. 在有足够企业财务数据的情况下，本文提出的测算企业贷款违约概率的方法，适用于计算企业的初始违约概率。商业银行可以通过本模型得到初始违约概率，并根据其划分一个初始信用等级，最后利用本课题组提出的贷款违约表法测算出最终违约概率[10]。这样得到的违约概率既考虑了客户的财务数据又考虑了客户的历史违约情况，大大提高了银行测算企业贷款违约概率的准确度，从而为本课题组提出的经济资本计量模型提供了数据基础[11]。

参考文献

[1] 中国银行业监督委员会.统一资本计量和资本标准的国际协议：修订框架.北京：中国金融出版社，2006年。
[2] 韩岗．国外信用风险度量方法及其适用性研究［J］．国际金融研究，2008 年第 3 期 ：43-48.
[3] 于立勇,詹捷辉．基于 Logistic 回归分析的违约概率预测研究财经研究［J］． 2004 年第 9 期：15-23.
[4] Laitinen,. Bankruptcy Prediction :Application of the Taylor Expansion in Logistic Regression[J]. International Reviews of Financial Analysis，2000 85—93
[5] 石晓军．边界 Logistic 违约率模型及实证研究［J］．管理科学学报，2007 年第 6 期:44-51.
[6] Ohlson. Financial ratios and the probabilistic prediction of bankruptcy [J] . Accounting Research , 1980 , 18 : 109—131.
[7] Zavgren C. Assessing the vulnerable to failure of American industrial firms : A logistic analysis[J] . Journal

of Business Finance and Accounting , 1985 , 12 : 1945.
[8] Lloyd.C. J. The use of smoothed ROC curves to summarize and compare diagnostic systems. Journal of the American Statistical Association, 1998，93, 1356–1364.
[9] 万柏坤,薛召,李佳等. 应用roc曲线优选模式分类算法［J］. 自然科学进展，2006年11期：1511-1516
[10] 彭建刚,易宇,李樟飞. 商业银行贷款违约概率测算方法探讨：贷款违约表法[EB/OL]. 中国科技论文在线，200804-202，http://www.paper.edu.cn.
[11] 彭建刚,张丽寒,刘波,屠海波. .聚合信用风险模型在我国商业银行应用的方法论探讨[J] . 金融研究，2008年第8期:72-84.

The research on the model of logistic default probability measure on the basis of the factor analysis

Peng Jiangang，Tu Haibo ，He Jing
College of Finance，Research Center of Financial Management ,Hunan University，Changsha，（410079）

Abstract

This paper proposes the logistic default probability measure model which is based on the factor analysis , in the light of the problem of missing original data 、multicollinearity and without time consideration in the general Logistic probability of default model. By introducing the factor analysis and calculating indices with time-weighted method, the paper improves the general Logistic default probability measure model, and then does the empirical research with the data of the Chinese Listed companies.The modle of logistic default probability measure based on the factor analysis can not only solve the problem of the data missing ,multicollinearity and Cramer, but also provides a relatively accurater result.

Keywords: *probability of default, factor Analysis, Logistic model*

作者简介：彭建刚，男，湖南长沙人，经济学博士，湖南大学研究院副院长，湖南大学金融学院教授、博士生导师，研究方向：金融管理与金融工程。
屠海波，男， 浙江杭州人, 湖南大学金融学院硕士研究生, 研究方向：金融管理与金融工程
何婧， 女， 湖南株洲人，湖南大学金融学院博士研究生，研究方向：金融管理与金融工程