企业经营管理预警:主成分分析在 logistic 回归方法中的应用

梁 琪
(南开大学经济学院, 天津 300071)

摘要: logistic 回归分析是度量企业信用风险的一种主流方法, 它的假设比较符合经济现实和金融数据分布的特点。但是考虑到现阶段我国上市公司的信用数据具有的高维性和高相关性等特点对 logistic 分析产生的负面影响, 本文在 logistic 分析中引入了能够有效降维和消除 logistic 方程共线性等问题的主成分分析, 并对我国沪深两市上市公司的经营失败进行了实证研究, 结果表明结合主成分分析法的 logistic 回归分析在模型解释和预测准确率等力面均优于简单的 logistic 分析。

关键词: 逻辑斯蒂分析; 主成分分析; 预警; 经营管理失败
中图分类号: F272.3 文献标识码: A 文章编号: 1004-6062(2005)01-0100-04

0 引言

Logistic 回归分析是量化度量企业信用风险的一种主流方法[3~6]。这种方法不仅本身灵活简便, 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的相关关系, 不要求变量服从协方差矩阵相等和残差项服从正态分布等, 这使得模型的分析结果比较客观[2,5]。但是, 我国上市公司信用数据具有高相关性和高维性(指标多)等特点[7]。这在应用 logistic 回归分析进行企业经营预警研究时会影响 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等问题。具体来说, logistic 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加。虽然系数自身的度量值不会由于共线性的存在而改变, 但系数的可靠性会因为其标准差的增加而降低, 从而使模型结果的稳定性, 特别是估计方程对新样本经营状况的预测准确率大幅度下降。另一方面, 在模型包含众多解释变量的情况下, logistic 分析的目标之一是如何得到预测企业经营状况的"节约模型"方程, 这个方程需要符合:①包含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具备经济意义上的说服力等条件[2]。一般来说, logistic 回归分析估计"节约模型"(Parsimonious Model)方程采用的方法多为逐步选择法, 即根据 logistic 方程似然比率统计量 G 的变化来选择最重要的解释变量。这种方法的缺点是它完全建立在统计方法的基础上, 没有考虑变量间的经济关系, 导致模型结果难以具有经济意义上的说服力, 而且估计方程的判别能力对模型原始样本具有很强的依附性, 但对测试样本或新样本的预测准确率则较差。此外, 这种方法还导致绝大部分的财务比率在逐选中被剔除掉了, 使得它们包含的企业信用风险信息也被排除在了估计方程之外。为了解决估计方程共线性和原始数据信息丢失等问题, 本文在 logistic 分析中引入了主成分分析法。它能够在模型具有众多解释变量的情况下, 同时实现降维和最大限度地减少原始数据中所含信息的丢失, 并且替代原始数据的主成分之间彼此互不相关[1,8]。主成分分析法的基本思路是:从 p 个相关的解释变量中推算出 k 个互不相关的主成分, 每一个主成分都是原始变量的线性拟合。第一个主成分最大限度地解释了原始变量数据的方差, 具有最大的特征值。第二个主成分与第一个主成分之间不存在线性相关关系, 它最大限度地解释了原始变量数据的剩余方差。以此类推, 第 k 个主成分与其他所有主成分之间均不存在线性相关关系, 它在所有主成分解释的原始变量数据的方差中排名第 k[1]。主成分之间互不相关的特点使得任意一个主成分的重要性可以通过它所解释的方差在所有主成分解释的方差中所占的比重来反映。如果前 k 个($k<p$)主成分可以解释大部分的原始变量数据的方差, 则 k 维主成分空间就能够最大限度地保留原始 p 维解释变量空间的信息。那么, 以这 k 个主成分作为 logistic 分析的解释变量来预测上市公司的经营失败, 就可以克服单纯的 logistic 分析存在的共线性和原始变量数据信息丢失等问题, 从而得到真正意义上的度量信用风险的"节约模型"。

文章以下的安排是这样的。第一部分对本文的研究方法进行了数学描述, 推导了结合主成分分析法的 logistic 回归分析模型;第二部分是实证分析;最后给出结论并总结全文。

1 研究方法

设模型样本数为 n, $X=(X_1, X_2, \cdots, X_p)^T$ 是一个 p 维随

收稿日期: 2003-04-09 修回日期: 2003-10-08
基金项目: 国家自然科学基金项目(70201001);国家社会科学基金项目(00CJY026)。
作者简介: 梁琪(1972—), 男, 陕西省蒲城县人, 南开大学金融学系副教授, 博士, 2001~2003 年赴日本一桥大学商学研究科做博士后研究, 现主要从事金融风险度量与管理、证券市场分析等领域的研究。

机向量, X 是企业财务比率, p 是选取的财务比率的个数。在财务比率取值范围不同或度量单位存在差异的情况下, 先对财务比率进行标准化处理。令标准化处理后的财务比率等于 W, 则

$$W_{m,i} = (X_{m,i} - \mu_{X_i}) / \sigma_{X_i}$$
$$m = 1, 2, K, n \quad i = 1, 2, K, p \tag{1}$$

主成分可以从经过标准化处理之后的财务比率的相关系数矩阵 η 的特征向量来推算, 令 λ 是相应的特征值, Π 是相应的特征向量, 则第 k 个主成分等于

$$Y_k = \Pi_k^T X = \pi_{1k} W_1 + \pi_{2k} W_2 + \Lambda\Lambda + \pi_{pk} W_p$$
$$k = 1, 2, K, p \tag{2}$$

Y_k 的方差以及 Y_k 和 Y_l 之间的协方差等于

$$Var(Y_k) = \Pi_k^T \sum \Pi_k$$
$$k = 1, 2, K, p \tag{3}$$

$$Cov(Y_k, Y_l) = \Pi_k^T \sum \Pi_l$$
$$l, k \in p, \ l \neq k \tag{4}$$

式(3)的求解等同于求拉格朗日函数表达式的最大化, 其推导过程为

$$\phi_k = \Pi_k^T \sum \Pi_k - \lambda_k (\Pi_k^T \Pi_k - 1)$$
$$k = 1, 2, K, p \tag{5}$$

对 Π_k 求偏导, 得

$$\partial \phi_k / \partial \Pi_k = 2 \sum \Pi_k - 2\lambda_k \Pi_k = 0$$

或

$$(\sum - \lambda_k I) \Pi_k = 0 \quad k = 1, 2, K, p \tag{6}$$

由于特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_k \cdots \geqslant \lambda_p$ 都是非负数, 因此特征向量可以通过式(6)计算得到, 再将其代入式(2)就可以求解主成分。在模型解释变量较多的情况下, 如果较少的 k 个主成分就能够解释原始变量数据大部分的方差, 那么这些主成分就可以在信息丢失最少的情况下实现对原有数据的替换, 并且它们之间是互不相关的。

在企业经营失败预测研究中, 作为两分类因变量的企业经营状况的概率取值在 0～1 之间, 这违背了回归分析的一些基本前提假设。但是通过对因变量进行 logit 转换, 使其取值区间在整个实数集中, 就可以采用 logistic 回归方法来研究企业经营状况与其财务比率之间的关系。因此, 企业经营失败分类和预测研究可以借助于一般 logistic 回归方程, 即

$$P_{a,t} = \frac{1}{1 + e^{-Z_{a,t}}} \tag{7}$$

其中 $P_{a,t}$ 是企业 a 在时期 t 内的条件违约概率, $Z_{a,t}$ 是企业 a 在时期 t 内的经营失败分类值, 后者可以通过以下这个多元回归方程来估计,

$$Z_{a,t} = \beta_{a,0} + \beta_{a,1} Y_{a,1,t} + \Lambda + \beta_{a,n} Y_{a,n,t} + e_{a,t} \tag{8}$$

其中 $Y_{a,1,t}, \cdots, Y_{a,n,t}$ 是企业 a 在时期 t 的经营失败解释变量, $\beta_{a,1}, \cdots, \beta_{a,n}$ 是解释变量的 logit 系数, $\beta_{a,0}$ 是常数项, $e_{a,t}$ 是误差项。

为了解决一般 logistic 模型中解释变量间存在相关关系从而影响模型分类和预测准确率的问题, 我们将以上主成分分析得到的 k 个主成分作为 logistic 回归分析的解释变量 $Y = (Y_1, Y_2, \cdots, Y_k)^T$, 带入 logistic 模型, 并令企业经营状况的条件概率为 $P(Z=1 \mid Y) = \pi(Y)$, Z 是企业的经营状况, $\beta = (\beta, 1, \beta_2, \cdots, \beta_k)^T$ 是解释变量(主成分)的 Logit 系数, β_0 是常数项, 则相应的 logistic 方程等于

$$\pi(Y) = \frac{e^{\beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + K + \beta_j Y_j}}{1 + e^{\beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + K + \beta_j Y_j}}$$
$$j = 1, 2, K, k \tag{9}$$

Logistic 回归分析采用最大似然法并通过似然函数来估计方程中解释变量的系数, 在二项 logistic 回归分析中, 似然函数等于

$$l(\beta) = \prod_{j=1}^{k} \pi(Y_j)^{z_j} [1 - \pi(Y_j)]^{1 - z_j}$$
$$j = 1, 2, K, k \tag{10}$$

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_0 和 β 的微分, 得到 $k+1$ 个似然方程式, 并令其等于 0, 即

$$\sum_{j=1}^{k} [z_j - \pi(Y_j)] = 0$$
$$j = 1, 2, K, k \tag{11}$$

和

$$\sum_{j=1}^{k} Y_{js} [Z_j - \pi(Y_j)] = 0$$
$$j, s = 1, 2, K, k \tag{12}$$

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　实证分析

2.1　模型样本

模型样本包括经营失败组和经营正常组。经营失败组选择了沪深两市 2000～2002 年间由于企业财务状况异常导致其 A 股股票交易被实行特别处理的上市公司 71 家①。这些企业来自中国证监会划分的 22 个上市公司行业部门类别中的 18 个, 具有很强的行业普遍性。因此在选择经营正常组的企业时, 模型只考虑了资产规模的因素, 并在所有非“ST”上市公司中随机抽选了 71 家企业, 模型样本总计为 142 家上市公司。本文还将模型样本随机划分为估计样本和测试样本。估计样本由 36 家“ST”和 36 家非“ST”企业组成, 用来估计 logistic 方程和检验分类准确率; 测试样本由 35 家“ST”和 35 家非“ST”企业组成, 主要用来检验预测准确率。

2.2　财务比率

财务比率的设计和选取是度量企业信用风险的出发点,

① 2000～2002 年间沪深两市共新增“ST”上市公司 93 家, 模型剔除了其中由于其他状况异常(前一期财务报告被会计师事务所出具无法表示意见或否定意见的审计报告等)导致“ST”的公司 22 家。

也是企业经营失败预测研究的关键。本文选取的财务比率数据大部分来自企业的财务报表，但也尝试纳入了部分能够反映投资者对企业未来经营前景预期的资本市场数据，如流通股股票的市场价格和流通股股本的数量等。在参考已有的研究文献以及考虑我国上市公司的特色和本研究需要的基础上，本文从能够反映企业赢利性、流动性、清偿性、增长性、景气和其它状况等六大类型指标入手，初选了 24 种企业财务比率(参见下表)。

表 1　模型最初选取的解释变量

财务比率名称	符号	财务比率名称	符号	财务比率名称	符号
景气分析		赢利性分析		增长性分析	
应收账款周转率	X_1	主营业务利润率	X_{10}	总资产增长率	X_{18}
其他应收款周转率	X_2	利息和税前利润率	X_{11}	净利润增长率	X_{19}
应付账款周转率	X_3	总资产利润率	X_{12}	其它	
固定资产周转率	X_4	总资产收益率	X_{13}	财务杠杆效应	X_{20}
总资产周转率	X_5	利息和税前收入/总资产	X_{14}	企业留存/总资产	X_{21}
流动性分析		股东权益收益率	X_{15}	股东权益比率	X_{22}
流动比率	X_7	清偿性分析		负债合计/资本	X_{23}
速动比率	X_7	资产负债率	X_{16}	流通股股本的市	X_{24}
现金比率	X_8	利息支付倍数	X_{17}	场价值/总负债	
平均营运资本/平均总资产	X_9				

2.3　一般 logistic 分析的结果

本文在引入主成分分析之前，首先对模型样本进行了简单的 logistic 分析，以利于后文的比较研究。模型采用了上市公司被特别处理前 3 年期的财务数据。由于上市公司在出现财务状况异常被“戴帽”之前，都已连续两年亏损或每股净资产低于股票面值，因此能否在前 3 年期时对企业的经营状况进行准确地预测就显得尤为重要。logistic 分析最终确定了 3 个最具有企业经营失败解释能力的变量——总资产利润率(X_{12})、负债合计/资本(X_{23})以及现金比率(X_8)，它们都与企业经营失败之间呈现负相关的关系，指标系数的 p 统计量均小于 0.005。而且与只有常数项的 logistic 方程相比，估计方程因变量方差中没有被解释的部分由 99.813 下降到了 34.004，方程的 Cox & Snell R^2 和 Nagelkerke R^2 统计量分别达到了 0.599 和 0.799，显示出了相当好的方程拟合度①。方程对于估计样本的分类准确率也很高，第一类错误仅为 8.3%，第二类错误为 11.1%，总分类准确率达到了 90.3%。尽管如此，对估计方程进一步的检验发现解释变量之间的相关程度较高，任意两个变量之间的相关系数均高于 0.50，最高达到了 0.81。Box-Tidwell 变换和正交多项(Orthogonal Polynomial Contrasts)对比等检验也显示出估计方程的解释变量之间存在着较高的线性函数关系，这些都违背了 logistic 回归分析的基本假设，估计方程共线性的问题降低了模型本身及其度量结果的可靠性。此外，模型估计的 logistic 方程只包含 3 个解释变量，这导致了绝大部分原始数据信息的丢失。测试样本预测准确率的明显下降证明了这些问题的存在，总预测准确率仅为 78.57%，第一类错误和第二类错误分别达到了 22.85%和 20.00%。

表 2　logistic 模型与结合主成分分析法的 logistic 模型的结果比较以及模型参数估计值

一般 logistic 模型		结合主成分分析法的 logistic 模型	
模型分析和检验结果比较			
−2LL	34.004	−2LL	50.275
Cox & Snell R^2	0.599	Cox & Snell R^2	0.497
Nagelkerke R^2	0.799	Nagelkerke R^2	0.663
总分类准确率	90.3%	总分类准确率	84.7%
总预测准确率	78.57%	总预测准确率	81.42%

模型的参数估计值

变量名称	代码	系数	标准差	变量名称	代码	系数	标准差
常数项		10.667	2.931	常数项		−0.001	0.355
总资产利润率	X_{12}	−99.242	28.221	盈利指标	F1	−4.285	1.101
负债合计/资本	X_{23}	−2.827	0.894	增长指标	F5	−1.073	0.514
现金比率	X_8	−5.220	1.968	资本市场指标	F3	−0.998	0.268

2.4　结合主成分分析法的 logistic 回归分析

企业财务比率的相关系数矩阵表明上市公司信用数据之间的高相关性主要来源于不同类型指标内部的比率之间、盈利性分析指标和增长性分析指标之间以及清偿性指标和其它指标之间，而流动性分析指标和景气指标与别的指标之间的相关性并不高。本文运用 SPSS 统计软件对被标准化处理之后的财务比率进行了主成分分析，并根据 Scree 检验，从 24 个主成分中抽取了 6 个，它们对原始财务比率方差的解释总计达到了 84.33%，这意味着它们包含了原始财务比率指标中的大部分信息。在综合考虑各个主成分解释企业经营失败的经济意义以及各种统计量的基础上，本文又从 6 个主成分中最终确定了 3 个作为 logistic 回归分析的解释变量，并依据各自的“因子负荷量”，即它们与原始财务比率之间的相关系数，分别定义它们为盈利指标、增长指标和资本市场指标。logistic 方程显示这些指标都与企业经营失败之间有着负相关的关系，指标系数的 p 统计量均小于 0.001。方程自身也显示出了较好的拟合度。与只有常数项的 logistic 方程相比，估计方程因变量方差中没有被解释的部分由 99.813 下降到了 50.275，卡方统计量为 49.538，显著性 p 值小于 0.001。方程的 Cox & Snell R^2 和 Nagelkerke R^2 计量分别达到了 0.497 和 0.663。方程对于估计样本的总分类准确率为 84.7%，其中第一类错误为 11.1%，第二类错误为 19.4%。尽管以上这些统计量和分类结果与简单的 logistic 回归分析相比有所逊色，但由于解释变量之间没有线性函数的关系，因此模型估计方程具有较高的可信度。对测试样本的检验进一步证明了引入主成分分析法之后的 logistic 回归方程在预测企业经营状况时呈现出的稳定性和可靠性，与分类检验

① 卡方统计量能够检验只包含常数项的 logistic 方程和模型估计的 logistic 方程的结果的比较是否显著。此处的卡方统计量等于 65.809，p 值小于 0.001。

结果相比，预测检验的准确率并没有出现明显的下降，总预测准确率达到了 81.42%，其中第一类错误为 17.14%，第二类错误为 20.00%，均优于简单的 logistic 回归分析。表 2 对一般 logistic 模型与结合主成分分析法的 logistic 模型的分析结果进行了比较，并给出了两个模型的参数估计值。

3 结束语

考虑到我国上市公司信用数据具有高相关性和高维性等特点，本文在 logistic 回归分析中引入了主成分分析法，它能够在解释变量众多的情况下，同时实现变量的降维和最大限度地减少变量中所含信息的丢失，并且通过这种方法得到的替代原始数据的主成分之间具有互不相关的关系，所有这些特点恰恰能够消除现阶段我国上市公司信用数据的分布特征对单纯 logistic 回归分析产生的负面影响。研究结果表明结合主成分分析法的 logistic 回归分析得到的企业信用风险度量方程更加稳定和可靠，对模型样本以外企业经营失败与否的预测准确率也更高。

参 考 文 献

[1] Basilevsky, A. Statistical factor analysis and related methods: theory and applications[M]. John Wiley & Sons, Inc. 1994.

[2] Hosmer, W. D., S. Lemeshow, Applied logistic regression[M]. John Wiley & Sons Inc. 2000.

[3] Huffman, P. S., J. D. Ward. The prediction of default for hight yield bond issues[J]. Review of Financial Economics 5(1996), pp. 75~89.

[4] Laitinen, K. E., T. Laitinen. Bankruptcy prediction: application of the taylor expansion in logistic regression[J]. International Review of Financial Analysis 9(2000), pp. 327~349.

[5] Liang, Q., Corporate financial distress diagnosis in China: empirical analysis using credit scoring models[J]. Hitotsubashi Journal of Commerce and Management Vol. 38, pp. 13~28.

[6] Martin, D. Early warning of bank failure[J]. Journal of Banking & Finance 1(1977), pp. 249~276.

[7] 王春峰，李汶华. 商业银行信用风险评估：投影寻踪判别分析模型[J]. 管理工程学报，2000(2).

[8] 张爱民，祝春山，许丹健. 上市公司财务失败的主成分预测模型及其实证研究[J]. 金融研究，2001(3).

Distress Prediction: Application of the PCA in Logistic Regression

LIANG Qi

(College of Economics, Nankai University, Tianjin 300071, China)

Abstract: Logistic regression model has become the standard method of analysis in the academic and practical fields of credit risk measurement. Considering the characteristics of high correlation and high dimension of the credit data of listed companies in China, this paper presents a new approach combining PCA with logistic analysis to empirically predict the Chinese corporate distress. The results show that the new approach is more stable and reliable than the simple logistic approach, especially in the credit prediction of the out of original samples.

Key words: Logistic Analysis; Principle Component Analysis; Credit Warning; Corporate Distress

责任编辑：杜 健