

边界 Logistic 违约率模型及实证研究^①石晓军¹, 任若恩¹, 肖远文²

(1. 北京航空航天大学经济管理学院, 北京 100083; 2. 北京华油天然气有限责任公司, 北京 100101)

摘要: 在前人研究基础上对边界 Logistic 违约率模型展开进一步研究. 首先通过抽样分布性质的研究从理论上说明了为什么边界 Logistic 违约率模型更优越. 然后利用中国公司数据展开实证研究, 不仅找到了 Cramer 问题的中国证据, 同时还发现边界 Logistic 违约率模型不仅能够克服 Cramer 问题, 而且对临界值不敏感、同时预测效率也相对较高.

关键词: Logistic; 违约; 边界 Logistic

中图分类号: F224.0

文献标识码: A

文章编号: 1007-9807(2007)03-0044-08

0 引言

利用 Logistic 方法建立违约率模型目前已经成为一种主流方法^[1-6]. 但最近, Cramer^[7]的研究表明, 一般的 Logistic 模型可能并不适用于贷款违约率的建模, 关键的原因是, 出现呆账这个事件本身并不服从 Logistic 分布. 他利用荷兰的商业银行经营的 627 笔呆账与 20 189 笔正常贷款构成的样本进行了实证研究, 给出的重要证据是, 一般的 Logistic 违约率模型难以通过 Hosmer-Lemeshow 拟合优度检验^②; 在 Hosmer-Lemeshow 检验中容易出现高估低端组的违约概率而低估高端组的违约概率的情况(本文称之为 Cramer 问题). 为此, Cramer 提出了边界 Logistic 方法. 尽管 Cramer 提出了边界 Logistic 模型, 但未对边界 Logistic 违约率模型比一般 Logistic 违约率模型优越的原因进行理论分析. 本文对边界 Logistic 违约率模型展开了进一步研究. 首先通过边界 Logistic 模型抽样分布性质的分析, 从理论上说明了为什么边界 Logistic 违约率模型更优越的原因. 然后利用一组中国上市公司的数据进行实证分析. 实证研究表明, 一般 Logistic 违约率模型除了容易出现 Cramer 的问题之外, 还

存在预测效率对临界点敏感的问题, 也就是, 一般 Logistic 违约率模型计算结果中的“灰色”区域较大, 不易确定分界点. 通过实证结果的对比发现, 边界 Logistic 违约率模型不仅能够克服 Cramer 问题, 而且对临界值不敏感, 同时模型的预测效率也相对较高.

1 边界 Logistic 与一般 Logistic 抽样分布性质的比较

Cramer 提出的边界 Logistic 模型的形式如下式:

$$P_i = \frac{\omega}{1 + \exp(-x_i^T \beta)} = \frac{\omega \cdot \exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \quad (1)$$

其中: ω 是在 $[0, 1]$ 之间的数, 由于 $\exp(-x_i^T \beta) > 0$, 因此 $P \leq \omega$, ω 实则是违约率的“边界”(bound). x_i 是第 i 个研究对象自变量向量的取值. β 是系数向量(包括截距 ξ).

下面先分析一般 Logistic 模型的任意配比的抽样分布.

根据一般 Logistic 模型的定义, 公司发生违约的概率为

$$P(y_i = 1 | x_i) = P^*(\theta, x_i) = P_i^* \quad (2)$$

① 收稿日期: 2005-05-27; 修订日期: 2007-04-02.

基金项目: 国家自然科学基金资助项目(70502005); 高校博士学科点基金资助项目(2005006004).

作者简介: 石晓军(1974—), 男, 江苏南通人, 博士, 副教授, Email: sxjstein@126.com.

② Hosmer-Lemeshow 拟合优度指标(通常简称为 H-L), 是由 Hosmer 和 Lemeshow 在 1989 年提出的一种 Logistic 模型拟合优度检验的方法. 关于 H-L 检验的具体内容, 可参考文献[8].

利用抽样样本就可以估计出参数 θ 。

假设使用的估计样本是从总样本中随机抽取的一个比率为 α 的抽样样本, 其中, 健全公司(即 $y_i = 0$ 的公司)的保留比率^③为 γ 。容易知道, 在随机抽取样本中, $y_i = 1$ 发生概率为^④

$$\alpha P_i^* \quad (3)$$

$y_i = 0$ 的发生概率为^⑤

$$\gamma\alpha(1 - P_i^*) \quad (4)$$

根据 Bayes 定律, 样本中 $y_i = 1$ 发生概率为^⑥

$$\tilde{P}_i = \frac{P_i^*}{P_i^* + \gamma(1 - P_i^*)} \quad (5)$$

由于抽取的比率 α 、 γ 是已知的, 把式(2)代入式(5), 这样就可以得到只关于 θ 的方程, 通过极大似然法就可以求出 θ 、 \tilde{P}_i 。

式(2)可以具体地写为

$$P_i^* = \frac{1}{1 + \exp(-x_i^T \beta)} = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \quad (6)$$

所以, 式(5)可以变为

$$\tilde{P}_i = \frac{\exp(x_i^T \beta)}{\gamma + \exp(x_i^T \beta)}$$

也即

$$\tilde{P}_i = \frac{\frac{1}{\gamma} \exp(x_i^T \beta)}{1 + \frac{1}{\gamma} \exp(x_i^T \beta)} = \frac{\exp(x_i^T \beta - \ln \gamma)}{1 + \exp(x_i^T \beta - \ln \gamma)} \quad (7)$$

从式(7)可以看出, 无论健全公司在抽样样本中的保留比率 γ 取什么值, 也就是说, 在构造样本时, 无论违约公司与健全公司的配比如何, \tilde{P}_i 总是遵从 Logistic 分布。

类似地, 由式(5)也可以求出采用边界 Logistic 模型时 $y_i = 1$ 发生概率为

$$\tilde{P}_i = \frac{\omega}{\omega(1 - \gamma) + \gamma + \exp(-x_i^T \beta + \ln \gamma)} \quad (8)$$

对比式(7)与(8), 可以看到, 采用一般

Logistic 方法时, 任意配比的抽样分布仍然服从 Logistic 分布; 采用边界 Logistic 方法时, 抽样分布是一个与配比方式有关的“广义”Logistic 分布。这个理论结果与已有的实证证据是一致的, Cramer 的研究^[7]表明, 样本的配比方式对 Logistic 违约率模型的估计有很大的影响, 一些配比比率下的 Logistic 违约率模型估计的拟合优度检验常不显著, 其根本原因是此时的抽样分布已经比较严重地偏离了一般 Logistic 分布, 可能服从一个广义的 Logistic 分布, 而边界 Logistic 违约率模型正好能够做到这一点。从这个意义上来说, 边界 Logistic 分布不仅仅是根据常识界定了违约可能的最大限度 ω , 同时还给出了抽样分布的广义 Logistic 形式, 使模型具有更大的灵活性、更强的适应性, 这正是边界 Logistic 违约率模型比一般 Logistic 违约率模型优越的根本原因。

2 边界 Logistic 模型的极大似然估计

本文采用最大似然估计法对边界 Logistic 模型中的参数进行估计。待估计的参数有: 截距项 ξ 、自变量的系数 $\beta_1 \cdots \beta_k$ 、“边界”变量 ω 。

容易求出样本容量为 n 时, 式(1)的对数似然函数

$$\begin{aligned} \ln[L(\theta)] &= \ln \left[\prod_{i=1}^n p_i^{y_i} (1 - p_i)^{(1 - y_i)} \right] \\ &= \sum_{i=1}^n [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \\ &= \sum_{i=1}^n \left[y_i \ln \left(\frac{p_i}{1 - p_i} \right) + \ln(1 - p_i) \right] \\ &= \sum_{i=1}^n \left[y_i (\ln \omega + \xi + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) + (1 - y_i) \ln(1 + (1 - \omega) \times \right. \\ &\quad \left. \exp(\xi + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})) - \ln(1 + \exp(\xi + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})) \right] \end{aligned}$$

③ 在 Logistic 违约模型的实证研究中, 存在着违约公司与健全公司最优配比的问题(请参考文献[9]), 由于实际经济生活中, 健全公司的数量通常很大, 因此在实证时要截出一部分, 只保留一部分健全公司。

④ 式(3)的含义: 如果没有从总样本中随机抽取的一个比率为 α 的抽样样本, 则 $y_i = 1$ 发生概率的概率是 P_i^* ; 如果从总样本中随机抽取的一个比率为 α 的抽样样本, 则该公司被抽取的概率为 α , 则在随机抽取样本中, $y_i = 1$ 发生概率的概率是 αP_i^* 。

⑤ 类似地, 在全部保留抽取的健全公司的情况下, $y_i = 0$ 发生概率的概率是 $\alpha(1 - P_i^*)$, 但是在实际研究时, 由于健全公司的数量通常很大, 只保留其中的一部分, 保留比率为 γ , 则在保留的抽样样本中, $y_i = 0$ 发生概率的概率即 $\gamma\alpha(1 - P_i^*)$, 此即式(4)。

⑥ 式(5)说的是在截去一部分健全公司的条件下, 保留样本中 $y_i = 1$ 发生的概率。假设总的随机抽取样本总量是 n , 则其中 $y_i = 1$ 的公司数为 $n\alpha P_i^*$, $y_i = 0$ 的公司数为 $n\alpha(1 - P_i^*)$, 又由于 $y_i = 0$ 的公司保留率只为 γ , 所以, 在最终保留的样本中, $y_i = 0$ 的公司数为 $n\gamma\alpha(1 - P_i^*)$, 因此, 在最终保留的样本中, $y_i = 1$ 发生的概率为 $n\alpha P_i^* / \{n\alpha P_i^* + n\gamma\alpha(1 - P_i^*)\}$, 也就是式(5)。(以上三个注脚得益于匿名审稿人的严格要求, 特致谢意!)

$$= \sum_{i=1}^n \begin{bmatrix} y_i(\ln\omega + \xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k) + (1-y_i)\ln(1+(1-\omega)\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)) - \ln(1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)) \end{bmatrix} \quad (9)$$

为了估计能使 $\ln[L(\theta)]$ 最大的总体参数 ξ , β 和 ω 的值, 可以分别对 ξ , β 和 ω 求偏导数, 然后令其等于零

$$\frac{\partial \ln[L(\theta)]}{\partial \xi} = \sum_{i=1}^n \left[y_i - \frac{\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} + \frac{(1-y_i)(1-\omega)\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} \right] = 0$$

$$\frac{\partial \ln[L(\theta)]}{\partial \beta_1} = \sum_{i=1}^n \left[y_i - \frac{\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} + \frac{(1-y_i)(1-\omega)\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} \right] x_1 = 0$$

$$\frac{\partial \ln[L(\theta)]}{\partial \beta_k} = \sum_{i=1}^n \left[y_i - \frac{\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} + \frac{(1-y_i)(1-\omega)\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} \right] x_k = 0$$

$$\frac{\partial \ln[L(\theta)]}{\partial \omega} = \sum_{i=1}^n \left[\frac{y_i}{\omega} - \frac{(1-y_i)\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + (1-\omega)\exp(\xi + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)} \right] = 0$$

这样, 就可以得到了一个非线性方程组, 利用 Newton-Raphson 迭代方法和 Mathematica 软件解出上述非线性方程组, 从而求出 ξ , β 和 ω 的值.

3 实证研究设计

3.1 样本

考虑到我国真正意义上的破产数据难以获得, 以及实证研究的可复制性, 本文采用国内研究的惯常做法, 以因财务状况异常而被特别处理 (ST) 作为上市公司陷入财务困境或违约的标志^⑦.

本文的样本来自 2002 年以前沪市上市的所有公司. 考虑到金融行业的特殊性, 本文的研究中不包括金融类上市公司. 一般地, 我国上市公司公布其当年年报的截止日期为下一年的 4 月 5 日, 故上市公司 $t-1$ 年的年报公布与其在 t 年是否被特别处理这两个事件是同年发生的. 考虑到以

上情况, 同时为了避免文献[2]所指出的高估模型预测能力问题, 本文采用的是上市公司 ($t-2$) 年的财务市场信息建立模型来预测其是否会在 t 年违约. 根据证券之星 (www.stockstar.com.cn) 的数据显示: 我国 A 股市场在 2003—2004 年 (截至 2004 年 8 月 9 日) 从沪市上市而被特别处理的公司一共有 52 家. 这 52 家指的是: 1) 2002 年没有被特别处理但 2003 年被特别处理的公司 20 家. 2) 2003 年没有被特别处理但 2004 年被特别处理的公司 32 家. 从这 52 家特别处理的公司中随机选取 2003 年被特别处理的 20 家及 2004 年被特别处理的 19 家作为训练样本中的违约公司. 把剩下的 2004 年被特别处理的 13 家公司作为检验样本中的违约公司. 而样本中的健康公司的配比比率采用 1:3, 也就是一个违约公司采用 3 个匹配的健康公司配比^⑧.

3.2 模型指标

参考文献[1, 3, 10, 11]等研究, 本文最初确定

⑦ 吕长江等(2004)对这个问题进行了概念的辨析, 尽管从概念辨析的角度是有道理的, 但从工具理性主义出发, 在当前我国的实际情况的约束下, 可能难以找到更好的替代方案, 因此本文将 ST、财务困境(危机)、违约视为同一概念. 需提醒的是, 文献[1]研究使用的是荷兰商业银行的呆账样本, 因此可以直接检验 Logistic 违约率模型的相关问题. 但本文采用的是我国上市公司的相关数据, 这主要是出于以下考虑: 第一, 缺乏公开可获取的、可信的贷款违约数据; 第二, 考虑到上市公司的财务数据通常是经过审计的, 相对未经审计的数据更为可信; 第三, 为了保证实证研究的可以重复性. 因此, 严格地说, 本文是使用我国上市公司的 ST 样本作为违约样本的替代样本(proxy). (感谢匿名审稿人的评论)

⑧ 文献[9]表明, 1:3 的配比比率比较适合中国情况的.

的备选指标共 24 个,通过指标的区别能力分析、多重共线性的剔除等指标遴选工作^⑨,最终进入模型的有:主营业务利润率、净资产收益率、每股收益盈利能力,销售净现率,每股现金净流量、利息保障倍数、现金债务比、资产留存收益的比率、资产负债率、Log(总资产)共 10 个指标。

3.3 临界点的选择

从理论上来说,临界点的选择对模型的判定效率有很大的影响.为了比较一般 Logistic 违约率模型与边界 Logistic 违约率模型对临界点的敏感性,本文设计了不同的临界点的情形。

首先,根据文献[12],Logistic 违约率模型的最优临界值公式为

$$p = \ln \frac{q_1 \cdot c_1}{q_2 \cdot c_2} \cdot \frac{1}{2} \quad (10)$$

其中: p 为违约临界值; q_1, q_2 为危机公司与非危机公司的先验概率; c_1, c_2 为 I 类错误和 II 类错误的成本。

q_1, q_2 可以根据我国上市公司当中危机公司所占比重大致求得,而 c_1, c_2 按照文献[13]的估计大致在 $[1/2, 1/38]$, 在研究中选取了 $1/2, 1/20$ 和 $1/38$ 三个具体数值。

根据我国的实际情况,可以大致估计出 q_2/q_1 , 如表 1。

表 1 2003—2004 年从沪市上市违约公司分布状况

Table 1 Distribution of defaulted listed companies in Shanghai stock market during 2003—2004

年份	违约公司数	上市公司总数	违约公司比率	非违约公司比率	q_2/q_1
2003	20	699	0.028 6	0.971 4	33.965
2004	32	699	0.045 8	0.954 2	20.834
总数	52	1 398	0.037 2	0.962 8	25.882

根据公式(10)可以计算出几种典型的违约的临界点取值如表 2。

表 2 几种典型的违约临界点取值

Table 2 Typical scenarios of the default cutoff-point values

	q_2/q_1	c_2/c_1	违约临界点取值
实际的典型情况	25.882	1/2	6.471
		1/20	0.647
		1/38	0.341
理论情况	1	1	0.5

注意到 6.471 的临界点取值没有意义,应舍弃,余下三种典型的违约临界点取值:0.647、0.341、0.5。

4 估计结果

4.1 一般 Logistic 违约率模型的估计与检验

一般 Logistic 违约率模型的估计结果与检验如表 3~表 5。

表 3 一般 Logistic 违约率模型回归系数及检验指标

Table 3 Regression coefficients and test statistics of the ordinary Logistic default model

变量	变量估计值	Wald 统计量	显著性水平 (sig.)
主营业务利润率	-13.706	4.135	0.042
净资产收益率	3.125	0.056	0.813
每股收益盈利能力	-2.813	2.277	0.131
销售净现率	-2.094	0.912	0.340
每股现金净流量	-0.102	0.092	0.761
利息保障倍数	-0.001	0.088	0.767
现金债务比	1.092	0.552	0.457
资产留存收益比率	-1.298	0.096	0.756
资产负债率	1.742	0.214	0.644
Log(总资产)	-0.597	0.788	0.375
常数项	13.182	0.948	0.330
-2log likelihood	56.668		
Cox&Snell R^2	0.533		
Nagelkerke R^2	0.789		

表 4 一般 Logistic 违约率模型的 Hosmer-Lemeshow 检验

Table 4 Results of Hosmer-Lemeshow tests of the ordinary Logistic default model

组别	0		1		总数
	观测值	预测值	观测值	预测值	
1	16	16.000	0	0.000	16
2	16	15.988	0	0.012	16
3	16	15.934	0	0.066	16
4	15	15.782	1	0.218	16
5	15	15.483	1	0.517	16
6	16	14.835	0	1.165	16
7	14	13.267	2	2.733	16
8	8	8.755	8	7.245	16
9	1	0.956	15	15.044	16
10	0	0.000	12	12.000	12
整体检验					
Chi-square	5.030		Sig.	0.754	

⑨ 限于篇幅,本文对指标遴选过程不做详细介绍,感兴趣的读者可向作者索要。

表 5 一般 Logistic 违约率模型的预测效率

Table 5 Prediction efficiency of the ordinary Logistic default model

分界点	总体误判率	I 类错误	II 类错误	I 类成本	II 类成本	总成本
0.5	0.269	0.154	0.385	1	1	0.269 5
0.647	0.115	0.154	0.077	20	1	0.150 3
0.341	0.231	0.077	0.462	38	1	0.086 9

注: ①总错判成本 = I 类错误概率 × 标准化的 I 类成本 + II 类错误概率 × 标准化的 II 类成本, 其中: 标准化的 I 类成本 = I 类成本 ÷ (I 类成本 + II 类成本); 标准化的 II 类成本 = II 类成本 ÷ (I 类成本 + II 类成本)
 ② I 类错误是指将违约公司误判为非违约公司, II 类错误是指将非违约公司误判为违约公司。

采用的检验样本^⑩是 2004 年 13 家未进入训练样本的违约公司和随机匹配的 13 家健康公司. 利用估计模型可以计算检验样本公司的违约概率, 其违约概率分布如图 1.

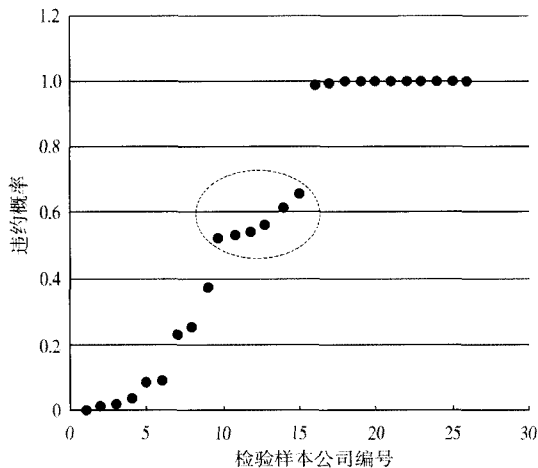


图 1 一般 Logistic 违约率模型的检验样本违约概率分布图

Fig. 1 Distribution of default probability in testing sample by the ordinary Logistic default model

4.2 边界 Logistic 违约率模型的估计与预测效率

利用 Mathematica 软件求解出边界 Logistic 的极大似然非线性方程组, 得到各参数的估计结果如下:

$$p = [0.9 \exp(40.82 - 163.8x_1 + 60.62x_2 - 18.22x_3 - 1.04x_4 - 1.596x_5 - 0.021x_6 + 9.9x_7 - 24.83x_8 + 3x_9 - 0.97x_{10})] / [1 + \exp(40.82 - 163.8x_1 + 60.62x_2 - 18.22x_3 - 1.04x_4 - 1.596x_5 - 0.021x_6 + 9.9x_7 - 24.83x_8 + 3x_9 - 0.97x_{10})]$$

其中: x_1, x_2, \dots, x_{10} 分别代表进入模型的 10 个变量.

按照以下方法对边界 Logistic 模型进行了 Hosmer-Lemeshow 检验.

首先, 按照边界 Logistic 模型估计结果由小到大进行排序. 然后分成规模大致相同的 10 个组 (第 10 组为 12 个, 其他均为 16 个). 第 3 步, 计算每组违约 (不违约) 的观察数与估计值. 第 4 步, 计算 Hosmer-Lemeshow 统计量:

$$G_{HL} = \sum_{i=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi_{10-2}^2$$

式中, O_j 第 j 组的观察值, $E_j = \sum_i \hat{p}_{ij}$ 第 j 组的估计值, n_j 第 j 组的样本点数. 该统计量服从自由度为 8 的卡方分布. 运算结果如表 6.

表 6 边界 Logistic 违约率模型的 Hosmer-Lemeshow 检验

Table 6 Results of Hosmer-Lemeshow tests of the bounded Logistic default model

组别	0		1		总数
	观测值	预测值	观测值	预测值	
1	16	16	0	0	16
2	16	15.977	0	0.023	16
3	16	15.913	0	0.087	16
4	15	15.168	1	0.832	16
5	15	15.093	1	0.907	16
6	15	14.790	1	1.210	16
7	15	14.235	1	1.765	16
8	8	8.902	8	7.098	16
9	1	0.977	15	15.023	16
10	0	0.000	12	12	12
整体检验					
Chi-square	3.414 3		Sig.		0.905

然后, 利用同一组检验样本来对模型进行检验, 其检验结果和 Logistic 违约率模型 (括号里面) 检验结果的对比如表 7.

表 7 边界 Logistic 违约率模型的预测效率

Table 7 Prediction efficiency of the bounded Logistic default model

分界点	总体误判率	I 类错误	II 类错误	I 类成本	II 类成本	总成本
0.5	0.192 (0.269)	0.077 (0.154)	0.307 (0.385)	1	1	0.192 0 (0.269 5)
0.647	0.192 (0.115)	0.077 (0.154)	0.307 (0.077)	20	1	0.088 0 (0.150 3)
0.341	0.192 (0.231)	0.077 (0.077)	0.307 (0.462)	38	1	0.082 9 (0.086 9)

⑩ 纵观现有的研究, 检验模型预测效力的方法大致可以分为四种: 构造模型样本的误分类率、其他样本的误分类率、交叉检验法、后续样本检验法. 从预测的本质看只有后续样本检验法才真正符合预测的应有之意. 因此, 本文主要采用后续样本检验法.

运用边界 Logistic 违约率模型计算出的检验样本的违约率分布如图 2 所示。

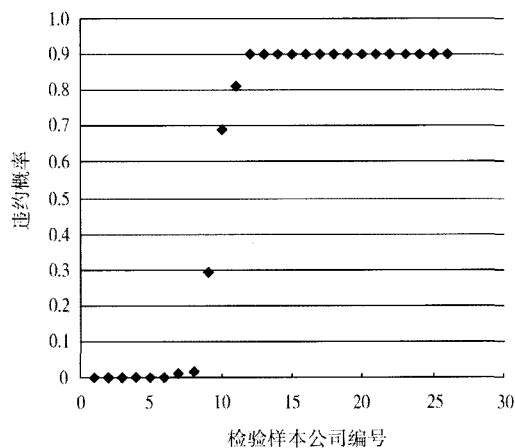


图 2 边界 Logistic 违约率模型的检验样本违约率分布图

Fig.2 Distribution of default probability in testing sample by the bounded logistic default model

5 结果分析

(1) 尽管一般 Logistic 违约率模型估计系数的显著性不理想, 但 R^2 的值比较大, H-L 检验的显著性高达 0.754, 因此不能拒绝模型拟合效果比较好的假设。也就是说, 一般 Logistic 模型的估计结果基本是可以接受的。

(2) 从一般 Logistic 违约率模型的 H-L 检验

来看, 也发现了 Cramer 指出的一般 Logistic 模型高估低端组违约率、低估了高端组违约率的现象。

首先, 从表 4 来看, 2、3 组是低端组。“0”列 (即正常组) 中, 这两组的预测值都比观察值低, 而“1”列 (即违约组) 中, 预测值都比观察值高。这表明一般 Logistic 模型高估了低端组违约率。同样地, 在表 4 中, 8、9 组是高端组。“0”列中, 8 组的预测值比观察值高 0.755, 9 组的预测值都比观察值 0.044, 总体上, 高端组的“0”列的预测值要比观察值高。类似地, “1”列中, 8 组的预测值比观察值低 0.755, 9 组的预测值比观察值高 0.044, 总体上, 高端组的“1”列的预测值要比观察值低。这表明一般 Logistic 模型低估了高端组违约率。

为了更充分地说明这个问题, 应该对低端组 (高端组) 估计违约率是否高于 (低于) 观察值进行严格的统计检验。将 Hosmer-Lemeshow 中的前 3 组定义为低端组, 而将后 3 组定义为高端组。仅用表 4 的结果来做这个检验, 显然样本数太少。为了解决这个问题, 首先从检验用的 13 个 ST 公司中随机地抽取 10 家公司, 将它们放入训练样本中, 进行一次 Hosmer-Lemeshow 检验分析, 照此, 共进行 99 次。这样, 可分别得到 300 组低端组和高端组。分别采用非参数的 Mann-Whitney 检验和参数的 t 检验对“低端组 (高端组) 估计违约率高于 (低于) 观察值”进行右侧 (左侧) 单边检验, H_0 假设为“低端组 (高端组) 估计违约率等于观察值”。检验结果如表 8 所示。

表 8 低端组 (高端组) 估计违约率高于 (低于) 观察值的 Mann-Whitney 检验和 t 检验

Table 8 Mann-Whitney test and t test of default probability in high-end and low-end groups

H ₀ 假设	低端组的检验		高端组的检验	
	观察值	估计值	观察值	估计值
分组	观察值	估计值	观察值	估计值
样本数量	300	300	300	300
均值	0.33	0.46	12.33	12.10
方差	0.33	0.62	16.33	19.97
M-W 统计量	86 901		88 387	
服从分布	N(90 150, 4 507 500)		N(90 150, 4 507 500)	
P 值	0.063		0.203	
t 统计量	2.24		0.681	
自由度	549		592	
P 值	0.016		0.253	

从表8的结果来看,至少有90%的把握说一般 Logistic 违约率模型高估低端组违约概率,因为 Mann-Whitney 检验和 t 检验的 P 值均小于 0.1.但大致只能在 0.25 的显著水平上接受一般 Logistic 违约率模型低估高端组违约概率的结论,因为 Mann-Whitney 检验和 t 检验的 P 值分别为 0.203 和 0.253. 总之,基本上可以认为在中国,也出现了一般 Logistic 违约率模型高估低端组违约概率、低估高端组违约概率的现象,即出现了 Cramer 问题.

(3)除 Cramer 问题之外,还发现了一般 Logistic 违约率模型的预测效果对临界点敏感,而边界 Logistic 违约率模型对临界点不敏感的证据.

从表5看到,对于一般 Logistic 违约率模型,当临界点选择为 0.647 时,总体误判率下降为 0.115,而临界点选择为 0.5、0.341 时误判率为 0.269 和 0.231. 这表明,一般 Logistic 违约率模型对临界点的选择十分敏感.图1也直观地说明了这个问题.在图1中,处于 0.5 附近的中间灰色地带的样本点比较多,如图中的虚线的椭圆所示.这部分样本点的归属对临界线的位置十分敏感.

从表7可以看到,边界 Logistic 违约率模型对临界值并不敏感,无论采用哪种临界值,边界 Logistic 违约率模型的总体错判概率、I 类错误和 II 类错误都是一样的.这从图2中也可以看出.在图2中,预测概率的分布比图1中的分布更加向两端分散,0.5 附近的中间灰色地带几乎被消除了,因此,边界 Logistic 违约率模型对临界值的选择并不敏感.这个性质决定了边界 Logistic 违约率模型比一般 Logistic 违约率模型更具有实际使用的前

景,它省去了通过错判概率和成本权衡临界点选择的复杂过程.

(4)边界 Logistic 违约率模型能比较有效地克服 Cramer 问题.

对于低端组,纠正高估违约率的倾向,意味着要调低违约率,也就是低端的预测违约率要向下分散.对于高端组,纠正低估违约率的倾向,意味着要调高违约率,也就是高端的预测违约率要向上分散.如果边界 Logistic 违约率模型能够克服 Cramer 问题,则它所估计的违约率分布图形就应该呈现出更加向两端分散的特点.对比图1和图2,发现图2表示的边界 Logistic 模型的预测违约率分布图相对于一般 Logistic 模型的预测违约率分布图更加向两端分散,这表明边界 Logistic 违约率模型能比较有效地克服 Cramer 问题.

(5)从错判概率和错判成本看,边界 Logistic 违约率模型也优于一般 Logistic 违约率模型.

一般地,犯 I 类错误的成本要比犯 II 类错误的成本要大.这样,模型预测能力的好坏主要是取决于其犯 I 类错误的大小.从表7可以看到,边界 Logistic 违约率模型犯 I 类错误的概率为 7.7%,而传统的 Logistic 违约率模型随着违约临界点的不同,犯 I 类错误的可能性分别为 15.4%、15.4% 和 7.7%.同时,从表7还可以看到,无论在哪种条件下,边界 Logistic 违约率模型的损失成本总比一般 Logistic 违约率模型小.

综上所述,边界 Logistic 违约率模型不仅能够克服一般 Logistic 违约率模型的问题,而且违约率预测的效率也比一般 Logistic 违约率模型好.

参 考 文 献:

- [1] Ohlson. Financial ratios and the probabilistic prediction of bankruptcy[J]. Accounting Research, 1980, 18: 109—131.
- [2] Zavgren C. Assessing the vulnerable to failure of American industrial firms: A logistic analysis[J]. Journal of Business Finance and Accounting, 1985, 12: 19—45.
- [3] 吴世农, 卢贤义. 我国上市公司财务困境的预测模型研究[J]. 经济研究, 2001, 6: 46—57.
Wu Shinong, Lu Xianyi. Research on financial distress forecasting model of China listed companies[J]. Economics Research, 2001, 6: 46—57. (in Chinese)
- [4] 马九杰, 郭宇辉, 朱 勇. 县域中小企业贷款违约行为与信用风险实证分析[J]. 管理世界, 2004, 5: 58—66
Ma Jiujie, Gou Yuhui, Zhu Yong. County SMEs' loans default behaviors and empirical analysis of its credit risks[J]. Management World, 2004, 5: 58—66. (in Chinese)
- [5] 管七海, 冯宗宪. 我国制造业企业短期贷款信用违约判别研究[J]. 经济科学, 2004, 5: 77—88
Guan Qihai, Feng Zhongxie. Discriminate analysis of China manufacturing companies' short-term loans default risk[J]. Economics

- Science, 2004, 5: 77—88. (in Chinese)
- [6] 梁 琪. 企业经营管理预警: 主成分分析在 logistic 回归方法的应用[J]. 管理工程学报, 2005, 1: 100—103
Liang Qi. Enterprise running and management warning: Applications of principal components logistic model[J]. Journal of Management Engineering, 2005, 1: 100—103. (in Chinese)
- [7] Cramer J S. Scoring banking loans that may go wrong—A case study [J]. Statistica Neerlandica, 2004, 58(3): 365—381.
- [8] 王济川, 郭志刚. Logistic 回归模型—方法及应用[M]. 北京: 高等教育出版社, 2001.
Wang Jichuan, Guo Zhigang. Logistic Regression Model: Methods and Applications[M]. Beijing: High Education Press, 2001. (in Chinese)
- [9] 石晓军, 肖远文, 任若恩. Logistic 违约率模型的最优样本配比与分界点研究[J]. 财经研究, 2005, 9: 39—49.
Shi Xiaojun, Xiao Yuanwen, Ren Ruoen. Optimal pairing and cutoff points of logistic default risk model[J]. Journal of Finance and Economics, 2005, 9: 39—49. (in Chinese)
- [10] Altman E I. Financial Ratios, Discriminate analysis and the prediction of corporate bankruptcy[J]. Journal of Finance, 1968, 23 (4): 589—609.
- [11] Shi Xiaojun. Robust factor credit discriminate model and empirical evidences from China[A]. Proceedings of the 7th International Conference on Industrial Management 2004[C]. Beijing: China Aviation Industry Press, 2004. 491—497.
- [12] 李皎予, 方军雄. 基于三因素模型的企业持续经营危机及其演化趋势的实证研究[R]. 北京: 湘财证券有限责任公司研究报告, 2003.
Li Jiaoyu, Fang Junxiong. Sustaining distress of enterprise based on three-factor-model and empirical study on its evolving trend. Beijing: Research Paper of XiangChai Security Co. Ltd., 2003. (in Chinese)
- [13] Altman E I, Haldeman R, Narayanan P. ZETA Analysis: A new model to identify bankruptcy risk of corporations[J]. Journal of Banking and Finance, 1977, 1: 29—54.

Bounded logistic default risk model and empirical evidences from China

SHI Xiao-jun¹, REN Ruo-en¹, XIAO Yuan-wen²

1. School of Management, Beijing University of Aeronautics and Astronautics, Beijing 100083, China;

2. Beijing HuaYou Natural Gas Co. Ltd. Beijing 100101, China

Abstract: Cramer pointed out shortcoming of plain Logistic default model and put forward bounded Logistic model. This paper does some further research about bounded Logistic default model. We first demonstrate why bounded Logistic default model is super to plain Logistic model theoretically through sampling distribution analysis. Then we give empirical evidences based on China companies' data. We not only find evidences about Cramer's problem, but also find that bounded Logistic model can solve the Cramer's problem, not sensitive to critical value and have higher prediction efficiency.

Key words: logistic; default; bounded logistic