文章编号: 1003- 207(2006) 04- 0025- 05

边界 Logistic 违约率模型 Bayes 分析及实证研究

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: Cramer[1] 指出了一般 Logistic 违约率模型容易出现的问题并提出了边界 Logistic 违约率模型。本文采用了不同于 Cramer(2004) 的 Bayes 分析方法对边界 Logistic 模型的后验分布的性质进行了分析, 从理论上说明了边界 Logistic 违约率模型更优越的原因。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 Bayes 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; Bayes 分析
中图分类号: C931 文献标识码: A

1 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[2~7]。大部分已有文献关注的重点是在不同的样本条件下, Logistic 违约率中具体指标的选择[2~6] 与整合[7]、Logistic 违约率模型与其他类型模型效率的比较[4],[6]、以及 Logistic 违约率模型本身效率的校验[2],[3]。而对 Logistic 违约率模型的一些基础性问题没有给予足够的关注。文献[8] 讨论了 Logistic 违约率模型的一类基础问题, 即最优样本配比与分界点问题。但是, 迄今为止, 对另一个更为重要的基础问题展开讨论的文献非常少见, 那就是 Logistic 违约率模型使用的前提, 即违约率的分布到底是否服从 Logistic 分布? 研究这个问题的困难之处在于, 违约率实际上是一种"隐变量", 我们在实际中无法直接观察得到, 因此, 我们也就不能采用简单的 Q- Q 图或其他的判断一个样本是否来自某类总体的统计检验方法。最近, Cramer (2004) 对这个基础性的问题进行了研究, 他利用荷兰的商业银行经营的 627 笔呆帐与 20189 笔正常贷款构成的样本进行了实证研究, 他的研究结果表明, 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, 出现呆帐这个事件本身并不服从 Logistic 分布。他给出的重要证据是, 一般的 Logistic 违约率模型难以通过 Hosmer-Lemeshow 拟合优度检验; 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。

本文的主要贡献在于采用了不同于 Cramer (2004) 的 Bayes 分析方法对边界 Logistic 模型的后验分布的性质进行了分析, 从理论上说明了边界 Logistic 违约率模型更优越的原因。并利用一组中国上市公司的数据进行实证分析。我们的实证研究表明, 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, 我们发现边界 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。

2 边界 Logistic 模型的 Bayes 分析

Cramer(2004) 提出的边界 Logistic 模型的形式如下式:

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

其中, ω 是在 $[0,1]$ 之间的数, 由于 $[1+\exp(-\mathbf{x}_i^T\beta)] \geqslant 1$, 因此 $\boldsymbol{P} \leqslant \omega$, ω 实则是违约率的"边界" (bound)。$\mathbf{x}_i$ 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)。

收稿日期: 2005- 07- 05; 修改日期: 2006- 07- 03
基金项目: 高校博士学科点基金资助项目(20050060004)
作者简介: 石晓军 (1974-), 男(汉族), 江苏南通人, 北京航空航天大学经济管理学院管理学博士, 副教授, 研究方向: 金融工程、信息产业经济学.

我们可以将(1)写成为式(2)的形式:

$$\text{logit}\Pr(y_i=1\mid x_i;\ \beta;\ \omega)=x_i^{T}\beta \qquad (2)$$

其中, $y_i=P_i/\omega$。

这个模型等价于:

$$y_i=1(z_i>0) \qquad (3)$$

其中, $1(\cdot)$ 是指示函数, 而 z_i 是一个连续的隐变量, 它表示违约的可能性, 它的分布取 logistic 概率密度的形式。

参考文献[8]的工作, 本文采用 t 分布近似法给出 Logistic 分布的概率密度函数如(4)式:

$$T_\upsilon(z\mid\mu,\sigma^2)=\frac{\Gamma((\upsilon+1)/2)}{\Gamma(\upsilon/2)\sqrt{\upsilon\pi}\upsilon}\left(1+\frac{(z-\mu)^2}{\upsilon^2}\right)^{-(\upsilon+1)/2} \qquad (4)$$

我们 ω 假设的先验分布来自正态分布, 如式(5):

$$P(\omega)=\frac{2}{\sqrt{2\pi}\sigma_\omega}e^{\frac{-(\omega-\mu_\omega)^2}{2\sigma_\omega^2}} \qquad (5)$$

由 Bayes 统计的分析方法我们可知, 在边界 Logistic 模型的条件下, z 的后验分布可表示为式(6):

$$T_\upsilon(z\mid\mu,\sigma^2,\omega)\propto\left(1+\frac{(z-\mu)^2}{\upsilon^2}\right)^{-(\upsilon+1)/2}\cdot e^{\frac{-(\omega-\mu_\omega)^2}{2\sigma_\omega^2}} \qquad (6)$$

我们假设 ω 的估计偏离 μ_ω 不总是很远, 则 $|-(\omega-\mu_\omega)^2/(2\sigma_\omega^2)|\approx 0$, 所以:

$$e^{\frac{-(\omega-\mu_\omega)^2}{2\sigma_\omega^2}}\approx 1+\frac{-(\omega-\mu_\omega)^2}{2\sigma_\omega^2} \qquad (7)$$

将(7)带入到(6), 我们可以近似地将 z 的后验分布看成为式(8):

$$T_\upsilon(z\mid\mu,\sigma^2,\omega)\propto\left(1+\frac{(z-\mu)^2}{\upsilon\sigma^2}\right)^{\eta} \qquad (8)$$

其中, $\eta<-(\upsilon+1)/2$。

也就是说 ω 先验信息的加入, 对 z 的分布形态并不会产生根本的改变, 但是它使得指数变得更小了, 从而使得后验分布的核比(4)式中的更为陡峭, 这相当于将一个比较"松弛"的 logistic 分布进一步地"压缩"为一个比较"紧凑"的 logistic 分布(也就是能使形态上较接近直线的"S"形向中间压缩, 在形态上更接近真正的"S"形)。这个性质有两点重要的含义, 一是边界 logistic 模型的效率要比较一般 logistic 模型高, 因为它使 logistic 分布形态的特征变得更加明显。二是中间的灰色区域被大大地压缩, 临界点附近难以正确判别所属类型的样本点将会减少。以上两点在本文后面的实证研究中均得到了支持。

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。

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。我们从这 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 家作为我们训练样本中的违约公司。把剩下的 2004 年被特别处理的 13 家公司作为我们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[9]。

3.2 模型指标

参考文献[2]、[4]、[9]、[10]、[11]等研究, 本文最初确定的备选指标共 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, 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, 在文献[12]的基础上可以得出 Logistic 违约率模型的最优临界值公式:

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

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]、[14] 的估计大致在[1/38, 1/2], 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

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. 0286	0. 9714	33. 965
2004	32	699	0. 0458	0. 9542	20. 834
总数	52	1398	0. 0372	0. 9628	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

	q_2/q_1	c_2/c_1	违约临界点取值
		1/2	6. 471
实际的典型情况	25. 882	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

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	. 042
净资产收益率	3. 125	. 056	. 813
每股收益盈利能力	– 2. 813	2. 277	. 131
销售净现率	– 2. 094	. 912	. 340
每股现金净流量	– . 102	. 092	. 761
利息保障倍数	– . 001	. 088	. 767
现金债务比	1. 092	. 552	. 457
资产留存收益比率	– 1. 298	. 096	. 756
资产负债率	1. 742	. 214	. 644
Log(总资产)	– . 597	. 788	. 375
常数项	13. 182	. 948	. 330
– 2log likelihood		56. 668	
Cox & Snell R^2		. 533	
Nagelkerke R^2		. 789	

我们采用的检验样本是 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。利用估计模型我们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:

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

组别	0		1		总数
	观测值	预测值	观测值	预测值	
1	16	16. 000	0	. 000	16
2	16	15. 988	0	. 012	16
3	16	15. 934	0	. 066	16
4	15	15. 782	1	. 218	16
5	15	15. 483	1	. 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

表 5　一般 Logistic 违约率模型的预测效率

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. 2695
0. 647	0. 115	0. 154	0. 077	20	1	0. 1503
0. 341	0. 231	0. 077	0. 462	38	1	0. 0869

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)

②I 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。

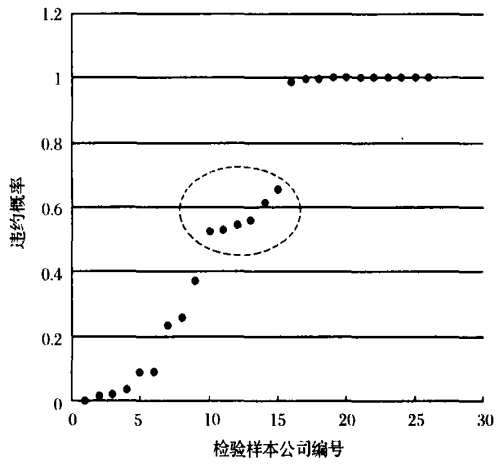

图 1　一般 Logistic 违约率模型的检验样本违约概率分布图

4. 2　边界 Logistic 违约率模型的估计与预测效率

我们利用 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:

$$p=\frac{0.9\exp(40.82-163.8x1+60.62x2-18.22x3-1.04x4-1.596x5-0.021x6+9.9x7-24.83x8+3x9-0.97x10)}{1+\exp(40.82-163.8x1+60.62x2-18.22x3-1.04x4-1.596x5-0.021x6+9.9x7-24.83x8+3x9-0.97x10)}$$

其中: x1, x2 …x10分别代表进入模型的 10 个变量。

然后, 我们利用同一组检验样本来对我们的模型进行检验, 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 6:

表 **6** 边界 **Logistic** 违约率模型的预测效率

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	0.1920 (0.2695)
0.647	0.192 (0.115)	0.077 (0.154)	0.307 (0.077)	20	1	0.0880 (0.1503)
0.341	0.192 (0.231)	0.077 (0.077)	0.307 (0.462)	38	1	0.0829 (0.0869)

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:

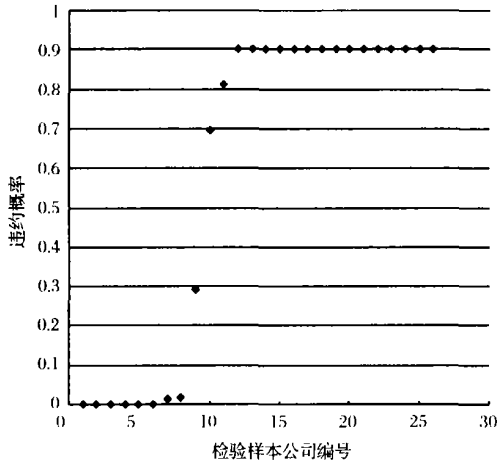

图 2 边界 Logistic 违约率模型的检验样本违约概率分布图

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, 因此我们不能拒绝模型拟合效果比较好的假设。也就是说, 一般 Logti 模型的估计结果基本是可以接受的。

(2) 从表 4 一般 Logistic 违约率模型的 H- L 检验来看, 我们也发现了 Cramer(2004) 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。

在表 4 中, 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。显然, 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 组的预测值都比观察值高 0.755, 9 组的预测值都比观察值低 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 组的预测值都比观察值低 0.755, 9 组的预测值都比观察值高 0.044, 总体上, 高端组的“1” 列的预测值要比观察值低。显然, 这表明一般 Logistic 模型低估了高端组违约率。

(3) 除 Cramer 问题之外, 我们还发现了一般 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 我们看到, 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 6 我们可以看到, 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, 我们发现图 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 6 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 6 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] Cramer. J. S.. Scoring Banking Loans that may go wrong – A Case Study [J]. Statistica Neerlandica, 2004, 58 (3): 365– 381.

[2] Ohlson.. Financial Ratios and the Probabilistic Prediction of Bankruptcy[J]. Accounting Research, 1980(18): 109– 131.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 梁琪. 企业经营管理预警: 主成分分析在 logistic 回归方法中的应用[J]. 管理工程学报, 2005, (1): 100– 103

[8] O_Brien SM, Dunson DB.. Bayesian multivariate logistic regression [Z]. 2004, http://citeseer.nj.nec.com/566151.html.

[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.

[10] Altman, Edward I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy[J]. Journal of Finance, 1968, 23(4): 589– 609.

[11] Shi xiao jun.. Robust Factor Credit Discriminate Model and Empirical Evidences from China[A]. Proceedings of the 7th International Conference on Industrial Management 2004, China Aviation Industry Press, 2004. 491– 497.

[12] Anderson, T. W.. An Introduction to Multivariate Statistical Analysis[M]. New York: Wiley, 1962.

[13] Altman, E., R. Haldeman, and P. Narayanan. ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations[J]. Journal of Banking and Finance, June, 1977.

[14] 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]. 湘财证券有限责任公司研究报告, 2003.

Bounded Logistic Default Model and Empirical Evidences from China

SHI Xiao jun[1], REN Ruo en[1], XIAO Yuan wen[2]

(1. School of Management, Beijing University of Aeronautics and Astronautics, Beijing 100083, China;
2. Beijing HuaYou Natrual Gas Co. Ltd., Beijing 100101, China)

Abstract: Cramer(2004) 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 superior to plain Logistic model theoretically through Bayes 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, which is not sensitive to critical value and has higher prediction efficiency.

Key words: logistic; default; bounded logistic; bayes analysis