Logistic 违约率模型的最优样本配比与分界点研究*

石晓军[1]，肖远文[2]，任若恩[1]
(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 风险管理部，北京 100101)

摘　要：Logistic 模型是研究违约率的主流方法之一，但目前的研究未能对最优样本配比与分界点这两个基本问题给予足够的重视。文章就此展开研究，设计了 15 种典型的样本配比—临界点的情景，通过实证比较的方法得出 1∶3 的样本配比与 0.647 的临界点比较适合我国的情况，而常用的 1∶1 的样本配比可能并不适用。

关键词：Logistic；违约；样本配比；临界点
中图分类号：F224.0　**文献标识码**：A　**文章编号**：1001-9952(2005)09-0038-11

一、引　言

Altman(1968)率先研究了公司破产预测的线性判别模型。从多元统计分析的结论我们知道，保证线性判别模型效率的两个前提，一是总体服从多元正态分布，一是协方差矩阵相等，这在现实经济中很难满足。于是，人们相继提出改进的办法，在这其中，Logistic 模型代表了一类重要的工作。中国的研究者自然对 Logistic 违约率模型在中国的实施与效率很感兴趣。迄今，已有相当多的学者对这个问题进行了研究，如吴世农等(2001)、马九杰等(2004)、于立勇等(2004)、梁琪(2005)等。但我们注意到，以上文献对 Logistic 模型在我国实施时的两个基本问题没有进行充分的研究，一是在我国建立 Logistic 模型时，样本中的违约公司与健康公司的配比是否会对模型的效率产生影响？二是适用于我国的 Logistic 违约率模型的最优分界点应该如何确定？本文拟对以上两个问题进行初步的研究。

收稿日期：2005-07-09

作者简介：石晓军(1974－)，男，江苏南通人，北京航空航天大学经济管理学院副教授；
肖远文(1878－)，女，江西九江人，北京华油天然气有限责任公司风险管理部；
任若恩(1948－)，男，北京人，北京航空航天大学经济管理学院教授，博士生导师。

二、文献综述

Logistic 分析方法假定企业发生违约的概率服从 Logistic 分布,采用一系列财务比率指标建立 Logistic 模型来预测企业发生财务危机的可能性,然后根据银行、投资者的风险偏好程度设定分界点,由此确定分析对象是否属于违约组。

较早采用 Logistic 模型对违约率进行分析的文献包括 Martin(1977)、Ohlson(1980)和 Zavgren(1985)。其中 Ohlson(1980)具有相当的代表性。他建立了一个 9 变量的多元 Logistic 违约率模型。他采用的破产样本来自 1970 年到 1976 年间的破产企业,而且在破产之前至少在证券交易所上市 3 年;配对样本是 2 058 家非破产的企业。Olhson 估计了三个模型,第一个模型预测一年内破产的企业,第二个模型预测第一年未破产而在第二年内破产的企业,第三个模型预测一年到二年内破产的企业。Ohlson(1980)特别分析了样本公司在破产概率区间上的分布以及两类判别错误和分界点的关系。他发现至少有 4 类显著影响公司破产概率的变量:公司规模、资本结构、业绩和当前资产的变现能力。特别需要指出的是,Ohlson(1980)提出为了避免"模型预测能力高估"的问题,应该采用违约前 2 年的数据建立模型。

Zavgren(1985)特别注意了样本配比问题,他采用了违约企业与非违约企业 1∶2 的配比方法,并指出 1∶2 的配比效果要比较 1∶1 好。但该文献并没有对样本配比问题进行系统的研究。

分界点的研究很早就得到重视,这是因为分界点的确定直接影响到 I 类错误和 II 类错误概率[①]。如果分界点定得过高的话,容易将本属于非违约的公司判为违约类公司,增加 II 类错误发生的可能性;但是如果违约分界点定得过低的话,容易将属于违约的公司判为非违约类公司,增加 I 类错误发生的可能性。理论上通常把判断违约的临界值设为 0.5。Anderson(1962)指出判别分析的临界值与两类错误的成本有关,他提出了一个确定最优临界值的公式。Theofanis(1987)证明了在破产公司分布概率一定的情况下预测模型最优临界概率取决于两类错误的分类的相对成本,随着 I 类错误成本上升,最优临界点随之下降,并在 Anderson(1962)的基础上提出了确定 Logistic 模型最优分界点的公式。

国内关于违约率模型研究的文献在不断地增长。如陈静(1999)、陈晓、陈治鸿(2000)、吴世农、卢贤义(2001)、李华中(2001)、于立勇(2004)、马九杰(2004)、姜天等(2004)、管七海等(2004)、梁琪(2005)等。国内的研究有几个基本的特点:第一,出于对国内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等考虑,国内通常以上市公司的财务数据为基础进行建模,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[②]。第二，国内的研究对哪些指标具有区别违约与非违约公司的显著能力给予了特别关注。陈静(1999)在单变量分析中，发现在负债比率、流动比率、总资产收益率、净资产收益率等 4 个指标中，流动比率和负债比率误判率最低；在多元线性判别分析中，由负债比率、净资产收益率、流动比率、营运资本/总资产、总资产周转率等 6 个指标构建的模型，在 ST 发生的前 3 年都能较好地预测 ST。陈晓、陈治鸿(2000)通过试验 1 260 种变量组合，发现负债权益比、应收账款周转率、主营利润率/总资产和预留收益/总资产对上市公司财务困境有着显著的预示效应。吴世农、卢贤义(2001)首先应用剖面分析和单变量判定分析，研究财务困境出现前 5 年内这两类公司 21 个财务指标各年的差异，最后确定了 6 个预测指标。第三，国内的研究在样本配比时几乎无一例外地采用了违约公司和非违约公司 1∶1 的结构。如，陈静(1999)以 1998 年的 27 家 ST 公司和 27 家非 ST 公司，使用了 1995～1997 年的财务报表数据构建样本。吴世农、卢贤义(2001)选取了 70 家 ST 公司和 70 家财务正常的公司构建样本。李华中(2001)选择 1997 年全部 ST 公司作为失败类组，选择一小部分 1999 年为 ST；而 1997 年、1998 年非 ST 类个股样本作为预测之用，再按照配对原则从同行业、相近资产规模的企业中选出同样数量的非 ST 公司作为非失败类组。最近的研究如管七海等(2004)、姜天等(2004)、梁琪(2005)等也是采用的 1∶1 的配比结构。只有于立勇等(2004)采用的是 35∶97、马九杰等(2004)采用的是 32∶43 的配比结构。以上研究情况表明，国内的研究没有意识到样本配比对模型效率影响的问题，基本默认 1∶1 的配比是最优的。实际上，由本文后面的内容可以知道这个配比可能并不是合适的。第四，国内的研究对表示模型效率的“误判率”给予了特别关注，如陈晓、陈治鸿(2000)报道的总体判别正确率为 78.24%；吴世农、卢贤义(2001)报道的前 1 年的误判率仅为 6.47%；前 4 年的误判率在 28%以内；李华中(2001)报道的模型预测误判率为 14.5%。由于各文献关于误判率的定义和时滞不同[③]，因此可比性并不强。

总的来说，国际上关于 Logistic 违约率模型的两个基本问题的研究已经有一些可借鉴的成果，但据我们所知，在实证层面详细展开这两个问题研究的文献并不多见。而国内的情况，据我们所知，迄今还没有文献专门讨论 Logistic 违约率的最优样本配比与分界点这两个基本问题。

下面，本文先从理论的角度证明任意匹配满足 Logistic 模型的必要条件，然后设计了几种典型的匹配比—分界点情景，采用实证比较的方法确定建立中国的 Logistic 违约率模型的最优样本匹配比与分界点。

三、任意配比的抽样分布

假设变量为 x_i，根据 Logistic 模型的定义，公司发生违约的概率为：

$$P(y_i=1|x_i)=P^*(\theta,x_i)=P_i^* \tag{1}$$

我们利用抽样样本就可以估计出 θ。

假设我们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）比率为 γ。容易知道，在随机抽取样本中，$y_i=1$ 的发生概率为：

$$\alpha P_i^* \tag{2}$$

$y_i=0$ 的发生概率为：

$$\gamma\alpha(1-P_i^*) \tag{3}$$

根据 Bayes 定律，样本中 $y_i=1$ 的发生概率为：

$$\tilde{P}_i=\frac{P_i^*}{P_i^*+\gamma\alpha(1-P_i^*)} \tag{4}$$

由于抽取的比率 α、γ 是已知的，把式(1)代入式(3)，这样我们就可以得到只关于 θ 的方程，通过最大似然法就可以求出 $\tilde{P}_i$。

式(1)可以具体地写为：

$$P_i^*=\frac{1}{1+\exp(-x_i^T\beta)}=\frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)} \tag{5}$$

所以，式(4)可以变为：$\tilde{P}_i=\dfrac{\exp(x_i^T\beta)}{\alpha\gamma+\exp(x_i^T\beta)}$

也即：

$$\tilde{P}_i=\frac{\frac{1}{\alpha\gamma}\exp(x_i^T\beta)}{1+\frac{1}{\alpha\gamma}\exp(x_i^T\beta)}=\frac{\exp(x_i^T\beta-\ln\alpha\gamma)}{1+\exp(x_i^T\beta-\ln\alpha\gamma)} \tag{6}$$

从式(6)我们可以看出，无论健全公司在抽样样本中的比例 γ 取什么值，也就是说，在构造样本时，无论财务危机的公司与健全公司的配比如何，$\tilde{P}_i$ 总是遵从 Logistic 分布。这个结论表明，我们可以以任意比率来配比财务危机公司和健全公司，都能够保证抽样分布满足 Logistic 分布，从而保证了 Logistic 分析方法的适用性。但是，这并不意味着在实证研究的过程中用任意的配比方法都是有效的。目前，国内的相关实证研究基本都采用 1∶1 的配比方法，Zavgren(1985)采用了 1∶2 的配比方法，并指出 1∶2 的配比效果要比 1∶1的好。那么，我们自然会问，什么样的配比是最适合我国的呢？下面我们将对这个问题展开详细的实证比较研究。

四、实证比较策略

1. 基本思路。为了确定适用于我国的最优样本配比与分界点，本文采取的基本策略是：首先设计出最可能出现的几种样本配比比率，并同时设计出常用的分界点，然后比较不同配比比率与不同分界点条件下 Logistic 模型的拟

合优度与后续样本预测效率，以此为主要依据判断适用于我国的最优样本配比与分界点。

2. 样本配比—分界点情景设计。违约临界值应该与发生的两类错误成本有关，Anderson(1962)给出了确定判别分析最优临界分数值的公式。Theofanis(1987)在 Anderson(1962)的基础上得出 Logistic 模型的最优临界值公式：

$$p=\ln\frac{q_1 c_1}{q_2 c_2}\cdot\frac{1}{2} \tag{7}$$

其中：p 为危机临界概率；q_1、q_2 分别为危机公司与非危机公司的先验概率；c_1、c_2 分别为 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中危机公司所占比重大致求得，而 c_1，c_2 则按照 Altman 等(1977)的估计大致在[1/2，1/38]之内，他们在研究中选取了 1/2、1/20 和 1/38 三个具体数值④。

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(7)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

情况/取值	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	—	—	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。

为了确定最优的样本配比比例，本文将选取几个典型的违约公司和健全公司的配对比率来作实证比较分析。所选的配对比率如表 3：

表 3　典型的配比比率与样本构成

比　率	违约公司	健全公司
1∶1	39	39
1∶2	39	78
1∶3	39	117
1∶5	39	195
全部健康公司	39	1 333

注：样本的选择方法可见下文。

下面要做的工作就是按照不同的配比比率确定的样本对 Logistic 模型的参数进行估计，然后根据不同的临界点，对模型的预测能力进行比较分析，根据比较的结果确定最优配比比率与临界点。

五、Logistic 模型的计量

1. 样本设计。本文样本来自 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 Ohlson(1980)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。

2. 模型的估计与评价。参考 Beaver(1968)、Altman(1968)、Ohlson(1980)、陈静(1999)、吴世农，卢贤义(2001)、Shi(2004)等的研究，本文设计了 24 个备选指标，在剔除多重共线性后，保留了对 ST 公司和非 ST 公司区分能力强的 10 个指标[5]：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)。待估计的 Logistic 模型就是：

$$p=\frac{1}{1+\exp(-(\alpha+\beta^{T}x))}$$

其中，$\alpha+\beta^{T}x=\alpha+x_1\beta_1+x_2\beta_2+\cdots\cdots+x_{10}\beta_{10}$；$x_1, x_2, \cdots\cdots, x_{10}$代表的是通过指标遴选过程而进入模型的 10 个变量。

关于 Logistic 模型的估计，本文采用最大似然估计法，具体的方法可以参考 Kleinbaum 等(1998)、王济川，郭志刚(2001)，或伍德里奇(2003)。对 Logistic 模型估计结果的评价，主要采用拟合优度和模型的预测能力两类指标。本文采用的是 Hosmer-Lemeshow 拟合优度指标[6]与后续样本预测能力指标。

六、实证结果

表 4 分别列出了配比比率为 1∶1、1∶2、1∶3、1∶5 和全部健康公司情况下的 Logistic 模型估计结果。

表 5 给出了不同配比比率情况下对估计模型进行的 Hosmer-Lemeshow 检验结果。

表 6 给出了不同配比比率、不同临界点的情况下，估计模型的后续样本的预测能力。

表 4　不同配比情况的 Logistic 模型的估计结果

配比比率	1∶1			1∶2			1∶3			1∶5			全部健康公司		
变量	变量估计值	Wald 统计量	显著性水平(sig.)	变量估计值	Wald 统计量	显著性水平(sig.)	变量估计值	Wald 统计量	显著性水平(sig.)	变量估计值	Wald 统计量	显著性水平(sig.)	变量估计值	Wald 统计量	显著性水平(sig.)
主营业务利润率	−1.208	1.375	0.241	−0.331	0.121	0.728	−13.706	4.135	0.042	−18.606	6.746	0.009	−0.305	0.419	0.517
净资产收益率	0.969	0.035	0.851	0.272	0.006	0.938	3.125	0.056	0.813	−10.981	0.352	0.553	−0.750	0.240	0.624
每股收益盈利能力	−1.327	1.436	0.231	−1.606	4.276	0.039	−2.813	5.277	0.022	−2.817	8.662	0.004	−0.751	10.957	0.001
销售净现率	1.677	4.256	0.040	0.207	0.048	0.827	−2.094	0.912	0.340	−1.774	0.903	0.342	0.016	0.014	0.907
每股现金净流量	−0.494	3.155	0.076	−0.307	5.783	0.015	−0.102	0.092	0.761	−0.221	0.258	0.611	−0.309	4.151	0.042
利息保障倍数	0.009	0.071	0.789	−0.049	8.397	0.005	−0.001	0.088	0.767	−0.007	0.146	0.702	0.002	0.526	0.468
现金债务比	−0.956	0.314	0.575	1.047	0.605	0.437	1.092	0.552	0.457	1.174	0.462	0.497	−0.112	0.011	0.915
资产留存收益比率	−1.806	0.317	0.574	−1.677	0.481	0.488	−1.298	0.096	0.756	−1.013	0.048	0.826	−3.604	4.428	0.035
资产负债率	0.170	0.003	0.955	0.897	0.147	0.701	1.742	0.214	0.644	4.799	4.487	0.033	0.031	0.000	0.984
Log(总资产)	0.670	10.988	0.000	0.017	0.003	0.958	−0.597	9.788	0.002	−0.296	0.258	0.612	−0.381	5.711	0.017
常数项	−13.501	2.013	0.156	−0.942	0.020	0.888	13.182	0.948	0.330	6.388	0.279	0.597	5.648	2.541	0.111
−2log likelihood	77.349			95.514			56.668			60.198			238.286		
Cox & Snell R^2	0.342			0.367			0.533			0.475			0.086		
Nagelkerke R^2	0.456			0.509			0.789			0.789			0.328		

表 5　不同配比情况的 Logistic 模型的 Hosmer-Lemeshow 检验结果

序号	1∶1					1∶2					1∶3					1∶5					全部健康公司				
	0		1			0		1			0		1			0		1			0		1		
	观测值	预测值	观测值	预测值	总数	观测值	预测值	观测值	预测值	总数	观测值	预测值	观测值	预测值	总数	观测值	预测值	观测值	预测值	总数	观测值	预测值	观测值	预测值	总数
1	7	7.494	1	0.506	8	11	11.782	1	0.218	12	16	16.000	0	0.000	16	23	23.000	0	0.000	23	112	111.905	0	0.095	112
2	7	6.858	1	1.142	8	12	11.329	0	0.671	12	16	15.988	0	0.012	16	23	22.994	0	0.006	23	112	111.619	0	0.381	112
3	7	6.288	1	1.712	8	10	10.635	2	1.365	12	16	15.934	0	0.066	16	23	22.969	0	0.031	23	111	111.314	1	0.686	112
4	5	5.233	3	2.767	8	12	10.086	0	1.914	12	15	15.782	1	0.218	16	22	22.912	1	0.088	23	112	110.985	0	1.015	112
5	8	4.678	0	3.322	8	11	9.626	1	2.374	12	15	15.483	1	0.517	16	22	22.814	1	0.186	23	111	110.659	1	1.341	112
6	3	4.172	5	3.828	8	8	8.854	4	3.146	12	16	14.835	0	1.165	16	23	22.507	0	0.493	23	109	110.217	3	1.783	112
7	3	3.070	5	4.930	8	8	7.870	4	4.130	12	14	13.267	2	2.733	16	23	21.925	0	1.075	23	111	109.634	1	2.366	112
8	0	2.142	8	5.858	8	4	5.490	8	6.510	12	8	8.755	8	7.245	16	23	20.879	0	2.121	23	109	108.784	3	3.216	112
9	0	0.910	8	7.090	8	1	2.166	11	9.834	12	1	0.956	15	15.044	16	13	14.618	10	8.382	23	108	106.942	4	5.058	112
10	1	0.154	7	7.846	8	1	0.162	8	8.838	9	0	0.000	12	12.000	12	0	0.382	27	26.618	27	85	87.941	26	23.059	111
χ^2	22.475					13.421					5.030					17.998					4.104				
Sig.	0.004					0.098					0.754					0.021					0.848				

表 6　模型的后续样本预测能力

配比比率	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
1∶1	0.5	0.154	0.154	0.154	1	1	0.1540
	0.647	0.078	0.308	0.104	20	1	0.2983
	0.341	0.269	0.077	0.462	38	1	0.0869
1∶2	0.5	0.154	0.308	0.091	1	1	0.1995
	0.647	0.231	0.462	0.124	20	1	0.4459
	0.341	0.162	0.154	0.231	38	1	0.1560
1∶3	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
1∶5	0.5	0.192	0.154	0.231	1	1	0.1925
	0.647	0.154	0.154	0.154	20	1	0.1540
	0.341	0.231	0.077	0.385	38	1	0.0849
全部	0.5	0.346	0.692	0.146	1	1	0.4190
	0.647	0.346	0.692	0.159	20	1	0.6666
	0.341	0.269	0.077	0.462	38	1	0.0869

注：(1)后续检验样本的构成方法：以保留的 2004 年 13 个未进入训练样本的违约公司为基础，然后再按照不同的配比比率随机选择健康公司，这样就可以构成不同配比比率条件下的后续检验样本。

(2)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)

七、实证结果分析

1. 样本配比比率与分界点对 Logistic 违约模型的估计与效率有明显的影响。我们从表 4 的结果可以看到，以不同的配比比率构成的训练样本，其 Logistic 模型参数估计结果差异十分明显，这说明样本的配比比率对最终的模型有着重要的影响，在实证研究中应谨慎选择样本配比比率。从表 6 的结果我们也可以看到，临界点对模型的误判率有着明显的影响，因此，在实证研究中也应谨慎选择临界点，这样才能保证模型的效率。

2. 健康公司的配比不应太大。为清晰起见，我们分别绘制了分界点为 0.341、0.5、0.647 时总体误判率、I 类错误与总成本随着样本配比比率变化的图形(如图 1)。该图比较清楚地揭示了两个规律：总体误判率、I 类错误与总成本在 1∶1、1∶2、1∶3 三种配比的情况下基本遵循先上升再下降的规律，这与 Zavgren(1985)的结论不太一致。该图也比较明显地揭示了“尾部上升”的规律，也就是当健康公司的配比太大时，模型的效率会出现下降的趋势。

3. 1∶1 的配比比率可能并不适合我国的情况。首先，根据表 5，从模型的

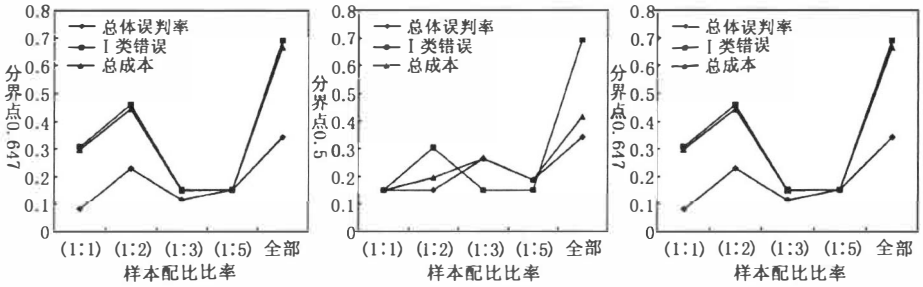

图 1　不同分界点条件下总体误判率、I 类错误与总成本随样本配比变化图

拟合优度 Hosmer 和 Lemeshow 指标来看，样本的配对比率为 1∶1、1∶5 时，其 χ^2 检验的显著性水平分别为 0.004、0.021，这表明我们能在 0.05 的显著水平上拒绝样本的配对比率为 1∶1、1∶5 时模型拟合效果可以接受的假设。尤其是配比比率为 1∶1 时，我们甚至能在 0.01 的显著水平上拒绝模型拟合效果可以接受的假设。同时，从该匹配比例情况下后续检验样本的违约概率分布图（如图 2）我们也可以看出，其违约概率分布基本可以看成是线性的，而不是 Logistic 分布。从本文的实证证据来看，1∶1 的配比比率实际上可能不太适合我国的实际情况。

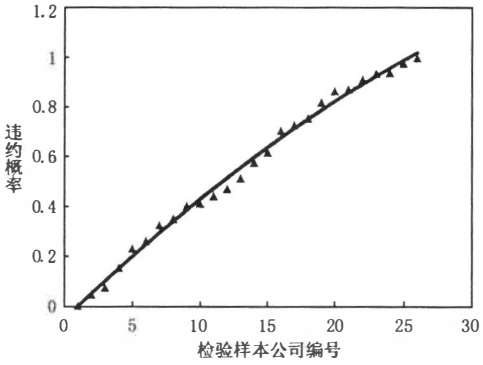

图 2　样本配比 1:1 时的后续检验样本违约概率分布图

4. 以 1∶3 为配比比率、0.647 为分界点比较适合我国的情况。为了确定出最优配比与临界点，本文采用以下方法：首先，我们设定凡误判概率（无论是 I 类还是 II 类误判概率）大于 35% 的情景均被剔除，余下的情景从误判概率的角度来说都是可接受的；接着根据总的误判成本最小的原则确定最优配比与临界点。根据第一条原则，可以剔除（1∶1，0.341）、（1∶2，0.647）、（1∶3，0.5）、（1∶3，0.341）、（1∶5，0.341）以及全部健康公司的所有情景。比较余下的 7 种情景，总体误判成本最小者是（1∶3，0.647）。因此，我们建议采用 1∶3 的样本配比结构和 0.647 的临界点来构建适合于我国的 Logistic 违约率模型。

* 本文是高校博士学科点基金项目《商业银行信贷资产风险中性定价与组合管理研究》成果之一。

注释：

①I 类错误是指将属于违约的公司误判为非违约公司，II 类错误是指将本属于非违约的公司误判为违约公司。

②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，因此本文也将 ST、财务困境(危机)、违约视为同一概念。

③综观现有的研究，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(Zmifewski，1984)。因此，本文主要采用后续样本检验法，即根据前几年的数据构建模型，然后用后来年份的数据进行检验。

④借鉴了李皎予，方军雄(2003)的研究。

⑤限于篇幅，本文正文部分不给出指标遴选的统计检验过程，如感兴趣可向作者索取。

⑥Hosmer-Lemeshow 拟合优度指标(通常简写为 HL)，是由 Hosmer 和 Lemeshow 在 1989 年研制出来的一种对 Logistic 模型进行拟合优度检验的方法。具体可参考 Kleinbaum 等(1998)或王济川，郭志刚(2001)。

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 dailure of American industrial firms: A logistic analysis[J]. Journal of Business Finance and Accounting,1985, (12):19～45.

[3]陈静. 上市公司财务恶化预警的实证分析[J]. 会计研究,1999,(4):31～38.

[4]陈晓，陈治鸿. 企业财务困境研究的理论、方法及应用[J]. 投资研究,2000,(6):29～33.

[5]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.

[6]李华中. 上市公司经营失败的预警系统研究[J]. 财经研究,2001,(10):58～64.

[7]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.

[8]Anderson, TW. An introduction to multivariate statistical analysis[M]. New York: Wiley, 1962.

[9]Cramer J S. Scoring banking loans that may go wrong — A case study [R]. Tinbergen Institute Discussion Paper, TI2000-090/4.

[10]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 1977,1(6): 29～54.

[11]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.

[12]于立勇，詹捷辉. 基于 Logistic 回归分析的违约概率预测研究[J]. 财经研究，2004,(9): 15～23.

[13]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.

[14]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.

[15]姜天，韩立岩. 基于 Logistic 模型的中国预亏上市公司财务困境预测[J]. 北京航空航天大学学报(社会科学版)，2004，(1)：54～58.

[16]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.

[17]Theofanis, T P. Corporate failure prediction models for the U. S. manufacturing and retailing sectors [D]. Ph. D. Dissertation of University of New York. University Microfilms Internationals, 1987.

[18]Shi xiao jun. Robust factor credit discriminate model and empirical evidences from China[C]. Proceedings of the 7th International Conference on Industrial Management 2004, China Aviation Industry Press, 2004, 491～497.

The Optimal Sample Pairing and the Critical Value of Logistic Default Risk Model: the China Case

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

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

Abstract: Logistic model is becoming more and more popular in default risk modeling. But the literatures by now have not paid enough attention to two kinds of fundamental questions concerning Logistic default modeling, that is, what is the optimal pairing structure of the training sample and what is the optimal cutoff point? This paper carries out a research into these problems. We design 15 typical scenarios of different sample pairing structure and cutoff point combination. Comparing the estimations of Logistic model and prediction efficiency in these different scenarios, it concludes that 1∶3 of sample pairing and 0.647 of cutoff value fit with China's data set, but the widely used 1∶1 structure may not be suitable for China case.

Key words: logistic; default; sample pairing structure; critical value

(责任编辑　许　柏)