基于 Lasso-logistic 模型的个人信用风险预警方法①

方匡南　章贵军　张惠颖

（厦门大学经济学院统计系）

【摘要】将 Lasso-logistic 模型引入个人信用评估，通过模拟实验发现，逐步回归法倾向于保留一些不重要的变量，而且选出正确模型的概率较低，而 Lasso 不仅计算更加快捷，可以同时进行变量选择和参数估计，而且能更准确地筛选出重要的变量。以信用卡消费信贷违约数据为例对我国个人信用评估进行实证分析发现，相对于全变量 Logistic 模型和逐步回归 Logistic 模型，Lasso-logistic 模型更能抓住影响消费信用风险的关键因素，而且预测准确率也更高。

关键词　信用风险　Lasso-logistic 模型　变量选择

中图分类号　F222　　文献标识码　A　　JEL 分类号　C52，D14

DOI:10.13653/j.cnki.jqte.2014.02.009

Individual Credit Risk Prediction Method: Application of a Lasso-logistic Model

Abstract: This article applies Lasso-logistic model to the individual credit risk prediction, the simulation suggests that stepwise method is apt to keep nonzero variable and the probability to select correct model is low, while Lasso has the advantages over other methods in computation, variable selection, parameter estimation simultaneously and grasping key factors. In additional, this article gives an empirical analysis of individual credit evaluation using credit cards data from a famous bank in China, finding that Lasso-logistic model relative to full logistic model and stepwise logistic model can select key factors and the predict accuracy is much higher.

Key words: Credit Risk; Lasso-logistic Model; Variable Selection

① 本文受到国家自然科学基金青年项目（71201139，71303200）、国家社科基金青年项目（13CTJ001）和教育部人文社科项目（12YJC790263）的资助。

引　言

随着金融创新和个人消费观念的改变，国内商业银行零售信贷业务，特别是个人信贷业务发展很快，其中，信用卡业务是发展最快的业务之一。据央行统计，2012 年信用卡消费金额占社会零售消费总额的 32%，信用卡发卡量持续增长，截至 2012 年底，信用卡发放量累计达到 3.31 亿张。随着信用卡业务规模的扩大，信用卡风险的防范与管理成为商业银行关注的焦点之一。信用卡的信用风险是指持卡人不能或不愿按照信贷协议约定偿还本息，从而对银行经营造成损失的可能性。个人信用风险是信用卡业务面临的主要风险。因此，构建个人信用评估系统，及时有效地应对可能发生的个人信用风险，不论从商业银行自身而言，还是从监管机构而言，都具有重大的现实意义。

个人信用风险评估的核心是建立针对不同客户类别的个人信用评分模型，根据信用评分模型对申请者的信用进行量化评分。发达国家的商业银行有着比较成熟的个人信用评估系统，但我国很多商业银行的个人信用评估体系尚在摸索阶段，即使已经构建的也存在不少问题。总的来说，我国的个人信用评估主要存在两大问题：首先，部分商业银行完全照搬发达国家的个人信用评估模型，但是由于国内外在消费观念和文化上的差异，因此，照搬国外的个人信用评估模型存在较多问题。其次，如何选取合适的评估指标体系。影响个人信用风险的指标很多，比如年龄、性别、婚姻、学历、家庭月收入等，而传统的 AIC、BIC 等子集选择法（Subset Selection）存在缺陷，主要因为：第一，子集选择法是一个离散而不稳定的过程，变量选择会受到数据集的微小变化而变化；第二，变量选择和参数估计分两步进行，后续的参数估计等没有考虑变量选择产生的偏误，从而低估实际方差；第三，子集选择法计算非常复杂（孙燕，2012）。本文将 Lasso-logistic 模型引入个人信用风险预警模型，科学地选择评估指标体系，以期构建合适我国国情且行之有效的个人信用风险评估模型，并提高个人信用风险的预警效果。

一、文献综述

从信用评分模型来看，主要方法有 Logistic 模型、判别分析方法、Bayes 方法、支持向量机、神经网络、Credit Metrics 模型和 KMV 模型等。胡心瀚等（2012）认为，最具代表性的 Logistic 模型由于预测准确率高、计算方法简单、变量解释能力强等特点被研究者广泛关注。方洪全和曾勇（2004）利用 Logistic 模型对商业银行实际数据分析后发现，Logistic 模型在金融机构对企业进行信用风险评估方面具有较强的预测能力。杨显爵等（2008）采用 Logistic 模型分析了台湾地区小额信贷违约问题。王来福和郭峰（2009）利用 Logistic 模型对中国住房抵押贷款信用风险问题进行了深入分析。平新乔等（2009）、葛君（2010）分别利用 Logistic 模型分析信用卡违约问题，均得到相对比较有说服力的分析结果。刘喜合和郭娜（2012）根据银行违约率微观数据，采用 Logistic 模型对我国住房抵押贷款信用风险的影响因素分析后表明，Logistic 模型的估计结果具有很强的稳定性。邓晶等（2013）选取 2010～2012 年 81 家 ST 公司 20 个财务指标数据，利用 Logistic 模型对我国上市公司的风险预测，结果表明该模型具有良好的预测效果。

此外，还有部分学者探讨其他信用风险评价方法，具有代表性的有：Wiginton（1980）比较了信用风险评估中两类常用的方法，Logistic 模型和判别分析，认为 Logistic 模型效果比判别分析相对要好一些。West（2000）的研究表明，虽然神经网络模型在信用风险预测

方面得到了较为广泛的应用，但神经网络模型与 Logistic 模型相比较并没有明显的优势。Lee 和 Sung（2000）比较 Logistic 模型和人工神经网络模型后认为，Logistic 模型鲁棒性较强，更适合分析城市消费信用违约问题。李志辉和李萌（2005）用不良贷款率作为信用风险高低的衡量标准，比较分析了主成分分析法、Fisher 线性方法、Logistic 模型、BP 神经网络方法对我国商业银行信用风险的识别情况，实证结果表明，Logistic 模型具有更强的信用风险识别和预测能力。韩岗（2008）比较分析了国内外各种应用于信用风险分析的模型后认为，Credit Metrics 模型由于需要可信的信用评级资料支持，以及操作性复杂、稳定性差等原因需要继续完善，而 KMV 对数据质量和数量要求比较高，应用到我国信用风险分析时机还不成熟，Logistic 模型在实际数据分析中拟合度高，比较符合我国实际情况，适合在实际应用中推广。

显然，利用 Logistic 模型分析我国信用风险问题具有较强的实用性。但是，以往对我国信用问题研究的对象主要是基于银行和上市公司数据，鲜有以个人信贷消费数据为分析对象研究我国信用风险问题，而后者相对前者不仅数据量庞大，而且所涉变量更多，研究难度也更大。一方面，包含过多变量的 Logistic 模型往往会因为多重共线性导致部分变量的检验统计不显著，这往往会降低模型的解释性并且影响模型的预测准确性；另一方面，个人信用风险评估模型一旦确定并选入一些无关的自变量，不仅会干扰对变量间关系的理解，而且会浪费人力物力搜集这些变量信息，甚至会给银行带来损失（王大荣等，2012）。因此，应用传统的全变量 Logistic 模型在分析以大规模消费信贷数据为对象的信用风险问题往往不适合，这时需要进行变量选择。

对于个人信用评分模型，变量选择是信用评分问题的重点和难点。以逐步（Stepwise）回归为代表的子集变量选择法，由于需要进行多次重复计算操作，算法复杂度是 np，当数据变量众多时，该方法往往就不适用了（Breiman，1995；孙燕，2012）。而影响个人信用风险的指标很多，因此有必要寻找合适的信用评分模型和变量选择方法，选取兼具解释性强和预测准确率高的方法。Lasso（Least Absolute Shrinkage and Selection Operator）方法由 Tibshirani（1996）提出，该算法作为一种变量选择和参数估计相结合的方法，由于其在大规模数据变量模型中具有良好的变量选择性质，受到广泛关注。Tibshirani（1996）指出，该方法兼具岭回归和子集选择的优点，计算简便，无需多次重复操作，最初广泛应用于生物学和医学领域。本文的主要贡献是通过模拟实验比较 Lasso、向前逐步回归和向后逐步回归法在自变量存在相关性时的优劣，并尝试将 Lasso-logistic 引入到个人信用预警中，构建适用于我国的个人信用评分模型。

二、Lasso-logistic 模型

1. Lasso 理论介绍

Tibshirani（1996）提出 Lasso 方法的动机来源于 Breiman（1995）的非负绞除法（Non-negative Garrote），该方法的目标函数可以概括为式（1）的形式。

$$\sum_{i=1}^{n}(y_i - a - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \qquad \text{s. t.} \quad c_j \geqslant 0 \quad \sum c_j \leqslant t \tag{1}$$

式（1）中非负因子 c_j 的加和受常数 t 控制，该算法思想是从最小二乘估计法出发，通过对部分非负因子 c_j 进行控制从而对估计的回归系数 $\hat{\beta}_j$ 进行压缩。Breiman（1995）用大量

数据模拟后得出结论：相对于子集选择方法，非负绞除法的预测误差相对较小，并且由于非负绞除法去除了模型中很多接近 0 但非 0 的变量，从而增强了模型的解释性；与岭回归方法比较，在高维数据分析中，非负绞除法由于对高维变量进行了压缩，模型简化了计算过程并且增强了重要变量的解释性。但是非负绞除法的缺点是其运算结果要依赖于最小二乘估计的符号和数值大小，并且在存在过度拟合和变量存在高度相关情况时，由于最小二乘估计效果不好而会影响预测准确性（Tibshirani，1996）。相比之下，Lasso 避免了非负绞除法的缺陷。

假设有数据变量 (X^i, y_i)，$i=1, 2, \cdots, n$，其中 $X^i=(x_{i1}, \cdots, x_{ip})$ 和 y_i 分别是解释变量和被解释变量的观测值。在一般的回归模型中，常常认为观测值彼此独立或者被解释变量 y_i 在给定解释变量 x_{ij} 的条件下相互独立。同时，假设 x_{ij} 是标准化的，即 $\frac{1}{n}\sum_i x_{ij}=0$，$\frac{1}{n}\sum_i x_{ij}^2=1$。令 $\hat{\beta}=(\hat{\beta}_1, \cdots, \hat{\beta}_p)^T$，Lasso 方法的估计量 $(\hat{a}, \hat{\beta})$ 定义为：

$$(\hat{a}, \hat{\beta}) = \text{argmin}\left\{\sum_i^n (y_i - a - \sum_j \hat{\beta}_j x_{ij})^2\right\} \quad \text{s. t.} \sum_j |\beta_j| \leqslant t \tag{2}$$

式（2）中 $t \geqslant 0$ 是调和参数，此时对一切的 t，有 a 的估计 $\hat{a}=\bar{y}$。为不失一般性，假定 $\bar{y}=0$，这样可将式（2）整理为式（3）。

$$\hat{\beta}=\text{argmin}\left\{\sum_i^n (y_i - \sum_j \hat{\beta}_j x_{ij})^2\right\} \quad \text{s. t.} \sum_j |\beta_j| \leqslant t \tag{3}$$

对调和参数 t 的控制将会使回归系数总体变小。若令 $\hat{\beta}_j^0$ 是回归参数的最小二乘估计值，$t_0=\sum_j |\hat{\beta}_j^0|$，$t<t_0$，就会使一些回归系数缩小并趋于 0，有些甚至会等于 0。例如，当 $t=t_0$，计算的结果就会使不为 0 的回归系数的数目减少到大约为 $p/2$ 个。上述式（3）的表达还可以用式（4）的惩罚函数的形式表达。

$$\hat{\beta}=\text{argmin}\left\{\sum_i^n [(y_i - \sum_j \hat{\beta}_j x_{ij})^2 + \lambda \sum_j |\beta_j|]\right\} \tag{4}$$

式（4）的第一部分表示模型拟合的优良性，第二部分表示参数的惩罚。调和系数 $\lambda \in [0, +\infty)$ 越小，模型的惩罚力度就越小，保留的变量就越多；λ 越大，模型的惩罚力度就越大，保留的变量就越少。子集变量选择方法，比如 AIC、BIC，是一个离散、无序的过程，变量或者被保留或者被删除，常常表现为高方差，不能降低整个模型的预测误差。而 Lasso 方法是一个连续的、有序的过程，方差较小。Tibshirani 提出的 Lasso 算法在模型变量选择时需要用二次规划方法求解。Efron 等（2004）认为，Tibshirani 的求解方法比较复杂，他们提出了计算速度更快的最小角回归算法（Least Angle Regression，LARS），并利用该算法计算 Lasso 参数路径（R 软件中有其对应的程序包 Lars）。

2. Lasso-logistic 模型

Lasso 方法主要应用于线性模型，其本质是在残差平方和上添加惩罚函数，在估计参数时，系数被压缩，部分系数甚至被压缩到 0 来实现模型选择。但是对于信用卡违约预测时，其因变量是二元离散取值，此时不能再利用线性回归模型，而应该使用 Lasso-logistic 回归。

假设有独立同分布的观测值 (X^i, y_i)，$i=1, 2, \cdots, n$，其中 $X^i=(x_{i1}, \cdots, x_{ip})$ 和

y_i 分别是模型的解释变量和被解释变量，并且 y_i 是二元离散数据变量，即 $y_i \in \{0, 1\}$，则 Logistic 线性回归模型的条件概率为：

$$\log\left\{\frac{p(y_i=1 \mid X^i)}{1-p(y_i=1 \mid X^i)}\right\}=\eta_\beta(X^i) \tag{5}$$

其中，$\eta_\beta(X^i)=\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j$。Lasso-logistic 回归模型中的系数估计值 $\hat{\beta}_\lambda$ 由式（6）凸函数的极小值给定：

$$S_\lambda(\beta)=-l(\beta)+\lambda\sum_{j=1}^{p}|\beta_j| \tag{6}$$

其中 $l(\cdot)$ 是对数似然函数，则式（6）中的 $l(\beta)$ 可以写成式（7）：

$$l(\beta)=\sum_{i=1}^{n}\{y_i\eta_\beta(X^i)-\log\{1+\exp[\eta_\beta(X^i)]\}\} \tag{7}$$

Lasso-logistic 回归模型中的系数估计值 $\hat{\beta}$ 可写成如式（8）的形式：

$$\hat{\beta}=\underset{\beta}{argmin}\sum_{i=1}^{n}\{y_i\eta_\beta(X^i)-\log\{1+\exp[\eta_\beta(X^i)]\}\}+\lambda\sum_{j=1}^{p}|\beta_j| \tag{8}$$

3. 调和参数 λ 的选择

Lasso-logistic 模型的变量选择，其关键在于调和参数 λ 的选取，常用方法方法有 Bootstrap、交叉验证、广义交叉验证法等，本文采用广义交叉核实（Generalized Cross-validation）方法确定罚参数 λ 的值，其具体算法如下：

若令 $p(\lambda)=\mathrm{tr}\{X(X^TX+\lambda(diag(|\hat{\beta}_1|, \cdots, |\hat{\beta}_p|))^{-1})^{-1}X^T\}$，由此可定义广义交叉验证值 GCV 统计量为式（9）所示形式：

$$GCV(\lambda)=\frac{\|y-X\beta(\lambda)\|^2}{n\{1-p(\lambda)/n\}^2} \tag{9}$$

显然，使交叉验证值 GCV 达到最小的罚函数为最优的罚参数 λ（Lambda），则最优 λ 的估计值为式（10）所示的表达式。

$$\hat{\lambda}=\mathrm{argmin}\ GCV(\lambda) \tag{10}$$

三、模拟实验

本文通过蒙特卡洛模拟方法比较 Lasso 变量选择法与向前逐步回归、向后逐步回归等传统子集选择法的优劣。数值分析的模型为：

$$\log\left\{\frac{p(Y=1 \mid X)}{1-p(Y=1 \mid X)}\right\}=\eta_\beta(X)=X^T\beta$$

其中，$\beta=(3, 1.5, 0.5, 0, 0, 0)$，$x_j \sim N(0, 1)$，且变量 x_i、x_j 间的相关系数为 $0.5^{|i-j|}$。我们分别模拟样本容量为 $n=50$、100、500，每种样本容量下重复 100 次实验。向前和向后逐步回归都基于 AIC 准则以及要求所选择的系数在 0.05 的显著性水平下显著。Lasso-logistic 基于 GCV 准则选择调和参数 λ。100 次模拟实验结果见表 1，其中“正确”列

表示 100 次实验中正确估计的零系数平均个数，即不重要变量正确剔除出去的平均个数。从表 1 可以看出 Lasso 方法估计正确的零系数个数要比向前和向后逐步回归更接近真实值，逐步回归法在变量选择过于保守，即倾向于保留一些不重要变量。“错误”列表示的是非零真实系数被错估为零的平均系数个数，即模型中重要变量被剔除出去的平均个数。从表 1 可以知道，Lasso 方法错估为零系数的平均个数要低于向前和向后逐步回归法。表 1 还给出 100 次模拟中选出正确模型的次数，即正确剔除不重要变量并保留重要变量的次数，可以看出 Lasso 方法选出正确模型的准确率远高于逐步回归方法。另外，图 1 给出 Lasso 方法估计的 Logistic 模型回归系数的 100 次模拟结果，可以看出回归系数结果接近参数真实值。

表 1　　100 次模拟实验结果

方　法	样本容量	零系数的平均个数		选择正确模型的次数
		正　确	错　误	
向前逐步回归	50	1. 69	0. 33	13
	100	1. 84	0. 17	16
	500	2. 09	0. 03	25
向后逐步回归	50	1. 71	0. 35	15
	100	1. 92	0. 27	16
	500	2. 11	0. 08	26
Lasso	50	2. 49	0. 31	66
	100	2. 73	0. 12	71
	500	2. 81	0. 02	88

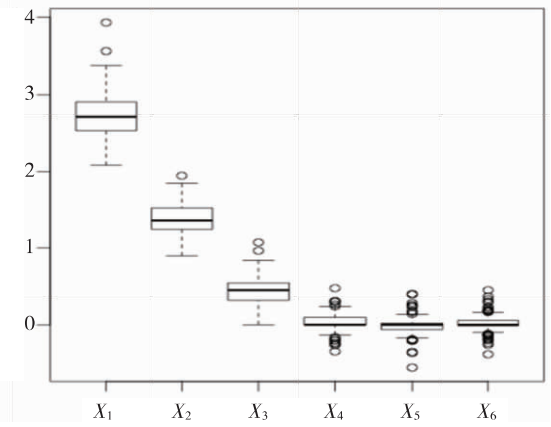

图 1　Lasso-logistic 模型 100 次模拟的系数箱线图

四、个人信用风险预警分析

1. 数据来源与变量说明

本文数据来自于中国某大型商业银行信用卡部，共收集到 131068 笔客户信用卡信贷数

据。数据变量如表 2 所示，共有 15 个方面的信息：客户是否为违约客户、信用卡张数、信用卡使用频率、户籍所在地、所在地都市化程度、性别、年龄、婚姻、学历、职业、个人月收入、个人月开销占家庭月收入的比例、家庭月收入、月刷卡金额、家庭人口数。

表 2　　变量说明

变　量	符　号	解　释
违约情况	Y	$Y=1$ 表示违约客户；$Y=0$ 表示非违约客户
信用卡张数	X_1	$X_1=1$ 表示 1 张；$X_1=2$ 表示 2 张；$X_1=3$ 表示 3 张；$X_1=4$ 表示 4 张；$X_1=5$ 表示大于 4 张
使用频率	D_1	$D_1=0$ 表示不经常使用；$D_1=1$ 表示经常使用
户籍所在地	D_2	$D_2=1$ 表示北部；$D_2=0$ 表示其他
	D_3	$D_3=1$ 表示中部；$D_3=0$ 表示其他
	D_4	$D_4=1$ 表示南部；$D_4=0$ 表示其他
都市化程度	D_5	$D_5=1$ 表示省会城市；$D_5=0$ 表示一般城镇
性别	D_6	$D_6=1$ 表示男；$D_6=0$ 表示女
年龄	X_2	$X_2\in[15,60]$；$X_2\in N$
婚姻	D_7	$D_7=1$ 表示未婚；$D_7=0$ 表示已婚
学历	D_8	$D_8=1$ 表示小学及以下；$D_8=0$ 表示其他
	D_9	$D_9=1$ 表示初中；$D_9=0$ 表示其他
	D_{10}	$D_{10}=1$ 表示高中；$D_{10}=0$ 表示其他
	D_{11}	$D_{11}=1$ 表示专科；$D_{11}=0$ 表示其他
职业	D_{12}	$D_{12}=1$ 表示无职业；$D_{12}=0$ 表示有职业
个人平均月收入	X_3	$X_3\in[0,+\infty)$
个人月开销占家庭月收入比例	X_4	$X_4\in[0,+\infty)$
家庭月收入	X_5	$X_5\in[0,+\infty)$
月刷卡额	X_6	$X_6\in[0,+\infty)$
家庭人口数	X_7	$X_7\in N$

注：N 表示整数集。

由于信用卡持卡用户是否会违约是重点研究问题，故将是否为违约客户的二元离散数据变量作为模型被解释变量，记为 Y。其余 14 个数据信息（商业银行认为可能会造成信用风险的信息）共转换为 19 个变量作为解释变量，其中部分数据变量属于二元离散数据变量，例如，使用频率、户籍所在地、都市化程度、性别等。

由于原始数据变量可能存在缺失、异常值等情形，所以在分析前有必要对缺失值进行填补与异常值的检测。由于本文缺失数据比例很低，因此以平均值法填补连续型数值变量的缺失值，以众数法填补二元离散数据变量的缺失值。考虑到解释变量中包含多个二元离散数据变量以及连续型数值变量，并且各变量单位各不相同，为了使模型参数估计系数具有可比

性，本文对连续型数值变量进行标准化处理。

同时，由于样本数据存在非对称性分布问题，即实际中非违约客户的数据数量远远大于违约客户数量，这会影响模型对数据资料相对较少的违约客户的预测精度。鉴于此，本文采用“减少多数法”对样本数据进行平衡处理，即运用抽样技术从样本较多的类别数据集中选取部分具有类别代表性的数据，用以降低数据变量类别间的不对称性。

为了比较模型的预测效果，本文在平衡前先将样本集分为训练集和测试集两部分。首先，从原始样本集中随机抽取约 2%的数据集，即 2533 笔数据作为测试集，留作用于对模型预测准确性的外推检验；然后，对剩余 128535 笔数据进行平衡处理，以 7051 笔违约用户数据为 1 单位，从非违约客户数据中随机抽取 7446 笔数据，使违约客户数据和非违约客户的比例大约平衡为 1：1，这样，训练集由 14497 笔数据构成。

2. 变量选择与模型估计

本文数据分析的 Lasso-logistic 模型利用 R 软件中 Glmnet 程序包，通过广义交叉验证，得到图 2 随着横坐标调和参数 λ 值的变化，纵坐标模型误差的变化情况，并在图 2 最上方给出模型筛选出来的对应变量数。图 2 中两条虚线中间的取值为 λ 正负标准差的值域范围，左边虚线表示使模型误差最小时的调和参数 λ 值取值。Tibshirani（1996）认为，λ 估计值取在此区间内模型预测偏差变动幅度相对较小，一般建议选取使模型相对简洁的 λ。此外，由于随着 λ 的取值不同，模型变量的压缩程度也会随之变化，即模型选择出的变量数目受 λ 估计值大小的影响。

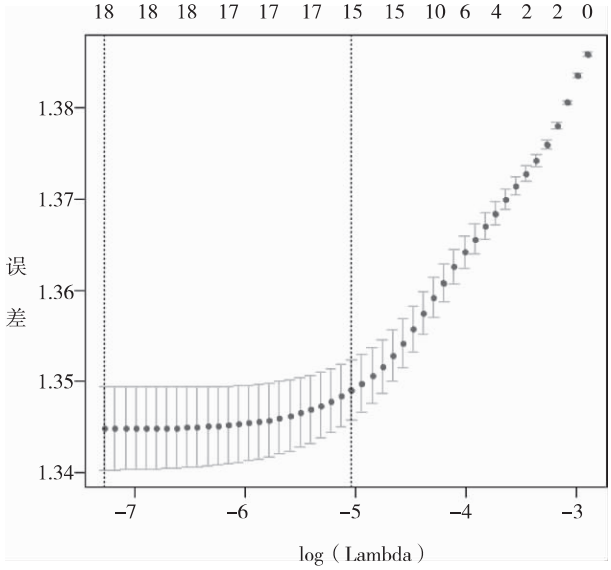

图 2 Lambda 与变量数目对应走势

图 3 显示随着调和参数 λ 值的变化，模型 19 个变量系数的筛选情况：随着 λ 取值变大，模型压缩程度增大，模型中包含的自变量个数减少，模型选择重要变量的功能增强。图 3 中如果取 $\lambda=e^{-3}$，那么对应挑选出来的变量是 D_2 和 D_{12}。为了尽量获得相对比较重要的变量，所以 λ 的理想取值应是使压缩程度达到最大，即获得的变量数目尽量少。根据 Tibshirani（1996）的取值经验，本文选取图 2 中贴近右虚线的 log（Lambda）值，即 $\lambda=e^{-5}$。此时，基于 Lasso 变量选择的 Logistic 模型参数估计结果如表 3 所示。

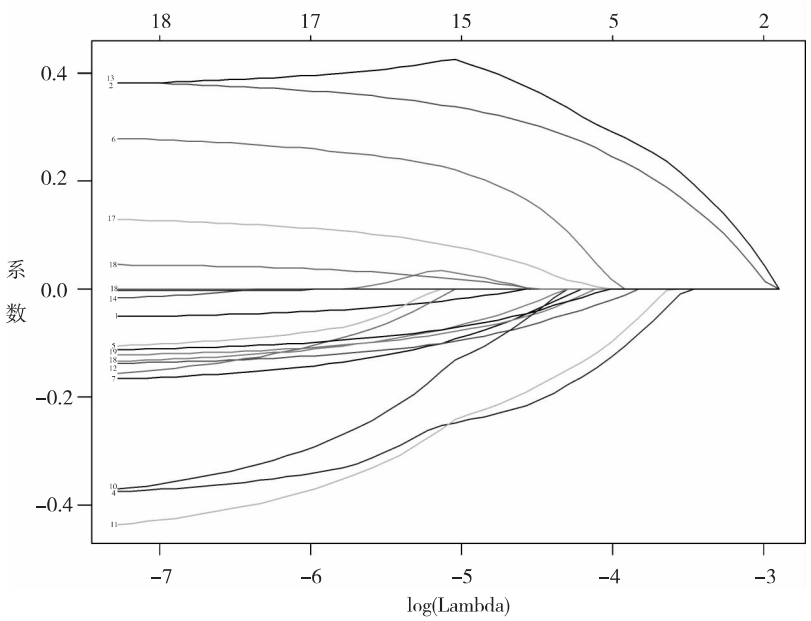

图 3 Lasso 系数解路径

表 3 列出全变量 Logistic 模型、Lasso-logistic 模型、向前逐步回归 Logistic 模型的估计结果，其中向后逐步回归法没有剔除变量，即向后逐步回归 Logistic 模型即是全变量 Logistic 模型，此处就不再列出向后逐步回归 Logistic 模型，下文的逐步回归 Logistic 模型均指向前逐步回归 Logistic 模型。从表 3 可以看出，对于全变量 Logistic 回归模型，变量 D_2、D_{12}和 X_4 前的系数均不显著，说明模型包含过多的解释变量；逐步回归 Logistic 模型剔除了不显著变量 D_2、D_{12}和 X_4，此时模型保留了 16 个变量；而 Lasso-logistic 模型剔除了变量 D_2、D_4、D_{12}和 X_4，选择出来的变量共 15 个，在 5%显著水平下这些变量均能通过检验。表明信用卡用户拥有的信用卡张数（X_1）、使用信用卡频率（D_1）、户籍所在地（D_2，D_3）、都市化程度（D_5）、性别（D_6）、年龄（X_2）、婚姻状况（D_7）、学历（D_8，D_9，D_{11}）、个人平均月收入（X_3）、家庭平均月收入（X_5）、月刷卡额（X_6）、家庭人口数（X_7）对用户的违约风险起到较为重要的影响作用。比较 Lasso-logistic 模型和逐步回归 Logistic 模型，Lasso 算法变量压缩效果更强，其筛选出的重要变量比逐步回归法少一个。因此，相对于其他两种 Logistic 模型而言，Lasso-logistic 模型更加简洁，更能抓住影响信用风险问题的关键因素。

表 3 参数估计

变　量	全变量 Logistic	Lasso-logistic	向前逐步回归 Logistic
截距项	−0.058	−0.230***	−0.098
X_1	−0.055***	−0.032***	−0.075***
D_1	0.390***	0.244***	0.382***
D_2	−0.056	0.050**	0
D_3	−0.437***	−0.189***	−0.388***

（续）

变　量	全变量 Logistic	Lasso-logistic	向前逐步回归 Logistic
D_4	−0.163**	0	−0.117***
D_5	0.286***	0.195***	0.270***
D_6	−0.174***	−0.108***	−0.180***
X_2	−0.140***	−0.090***	−0.141***
D_7	−0.126***	−0.075***	−0.120***
D_8	−0.404***	−0.167***	−0.393***
D_9	−0.462***	−0.213***	−0.463***
D_{10}	−0.179***	0	−0.175***
D_{11}	0.369***	0.317***	0.376***
D_{12}	−0.026	0	−0.128***
X_3	−0.141***	−0.088***	0.129***
X_4	−0.001	0	0
X_5	0.135***	0.088***	0.052***
X_6	0.049***	0.029***	−0.098***
X_7	−0.116***	−0.072***	0

注：***、**和* 分别表示在 1%、5%和 10%的水平下显著；系数为 0 表示在变量选择中，被剔除出去的变量。

从 Lasso-logistic 模型的结果可以看出，违约风险相对较高的信用卡持卡用户居住在北部地区省会城市、个人月收入相对较低（1000 元以下）、家庭月收入相对较高（8000 元以上）、使用信用卡频率较高及月刷卡额度较高（20000 元以上）、年龄相对较小（35 岁以下）且学历为专科的已婚女性。本文的分析结果与 Carow 和 Staten（1999）的研究结论在一定程度上接近，而与 Schreiner（2004）、葛君（2010）的研究结果存在明显差异。Carow 和 Staten（1999）研究认为，信用卡的使用者往往是年龄较小、受教育程度较高以及有较多信用卡的人群信用违约风险往往较高；Schreiner（2004）分析玻利维亚商业银行的数据后认为，信用风险存在性别差异，并且女性的信用违约率低于男性；葛君（2010）利用因子分析方法结合 Logistic 模型对中国新余市某商业银行信用卡客户数据进行分析后认为，具有高中学历、年龄小于 25 岁、未婚、初级职称的客户违约概率大。

3. 模型准确率比较

接下来，我们分别测试上文建立的全变量 Logistic 模型、逐步回归 Logistic 模型以及 Lasso-logistic 模型的预测准确率。表 4 给出这 3 个模型分别在训练集和测试集中对违约客户和非违约客户预测准确率对比情况。

表 4 模型预测准确率比较 (单位:%)

模 型	训练集准确率			测试集准确率		
	违约	非违约	总体	违约	非违约	总体
全变量 Logistic	57.22	64.17	60.79	56.90	62.86	62.09
逐步回归 Logistic	64.45	66.82	65.67	61.78	65.66	65.16
Lasso-logistic	73.42	79.36	76.47	70.69	75.25	74.66

注:测试集中数据是通过对测试集进行 50 次随机取样后计算所得平均值。

在实际的信用风险评价中,将违约客户误判为非违约客户对授信人或社会而言造成的潜在经济损失会更大。因此,对违约客户的准确判断更为重要,其次考虑的才是对非违约客户的判断。表 4 计算结果表明:无论是在训练集还是在测试集中,Lasso-logistic 模型对违约客户预测的准确率是 3 个模型中最高的,其中训练集中,Lasso-logistic 模型比全变量 Logistic 模型和逐步回归 Logistic 模型分别高出 16 和 9 个百分点;在训练集中,Lasso-logistic 模型对非违约客户预测准确率也是最高的,比全变量 Logistic 模型和逐步回归 Logistic 模型分别高出 15 和 12 个百分点;在测试集中,Lasso-logistic 模型对违约客户和非违约客户的准确率也是最高的,而且总体的预测准确率维持在 74.66%,说明具有较好的外推性。

五、结 论

本文将 Lasso-logistic 引入到个人信用风险评估中来,首先通过模拟实验分析比较了在当自变量存在着较强相关性下,Lasso、向前逐步回归和向后逐步回归法在自变量存在相关性时的优劣,并以我国某大型商业银行的信用卡数据为例对个人信用评估进行了实证分析,本文主要结论有:第一,模拟实现发现,向前逐步回归和向后逐步回归在变量选择时倾向于保留一些不重要的变量,也就是说所选出的模型相对复杂,难以剔除一些对因变量不重要的自变量,而且选出正确模型的概率较低,而 Lasso 变量选择方法相对于向前逐步回归和向后逐步回归,不仅计算更加快捷,能够同时进行变量选择和参数估计,而且能更准确筛选出相对重要的变量;第二,全变量 Logistic 模型将所有变量选入模型,然而并不是所有变量参数均能通过显著性水平检验,这在一定程度上降低了模型解释性,Lasso-logistic 模型和逐步回归 Logistic 模型克服了全变量 Logistic 模型多重共线性的同时也增强了模型解释性,相比逐步回归方法而言,Lasso 算法压缩效果更好;第三,在训练集中,Lasso-logistic 模型不管是对违约客户的预测还是非违约客户的预测,其准确率都是最高的,并且对测试集的预测也保持了较高的准确率,说明 Lasso-logistic 模型具有较高的外推性。

因此,将 Lasso-logistic 模型引入个人信用风险预警模型,可以更加科学地选择评估指标体系,并构建合适我国国情且行之有效的个人信用风险评估模型,提高个人信用风险的预警效果。

参 考 文 献

[1] Efron B., Hastie T., Johnstone I., Tibshirani R., 2004, *Least Angle Regression* [J], Annals of Statistics, 2 (32), 407~499.

[2] Breiman L., 1995, *Better Subset Regression Using the Nonnegative Garrote* [J], Technometrics, 4

(37), 373～384.

[3] Carow K. A., Staten M. E., 1999, *Debit, Credit, or Cash: Survey Evidence on Gasoline Purchases* [J], Journal of Economics and Business, 5 (51), 409～421.

[4] Lee T. H., Jung S. C., 1999, *Forecasting Credit Worthiness: Logistic vs. Artificial Neural Net* [J], The Journal of Business Forecasting Methods & Systems, 4 (18), 28～30.

[5] Schreiner M., 2004, *Scoring Arrears at a Aicrolender in Bolivia* [J], Journal of Microfinance, 2 (6), 65～88.

[6] Tibshirani R., 1996, *Regression Shrinkage and Selection via the Lasso* [J], Journal of the Royal Statistical Society, Series B, 1 (58), 267～288.

[7] West D., 2000, *Neural Network Credit Scoring Models* [J], Computers & Operational Research, 11 (27), 1131～1152.

[8] Wiginton J. C., 1980, *A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior* [J], Journal of Financial and Quantitative Analysis, 3 (15), 757～770.

[9] 邓晶、秦涛、黄珊：《基于 Logistic 模型的我国上市公司信用风险预警研究》[J]，《金融理论与实践》2013 年第 2 期。

[10] 方洪全、曾勇：《银行信用风险评估方法实证研究及比较分析》[J]，《金融研究》2004 年第 1 期。

[11] 葛君：《基于 logistic 模型的信用卡信用风险研究》[J]，《中国信用卡》2010 年第 12 期。

[12] 韩岗：《国外信用风险度量方法及其实用性研究》[J]，《国际金融研究》2008 年第 3 期。

[13] 胡心瀚、叶五一、缪柏其：《上市公司信用风险分析模型中的变量选择》[J]，《数理统计与管理》2012 年第 6 期。

[14] 李志辉、李萌：《我国商业银行信用风险识别模型及其实证研究》[J]，《经济科学》2005 年第 5 期。

[15] 刘喜合、郭娜：《我国住房抵押贷款信用风险因素分析》[J]，《山东社会科学》2012 年第 3 期。

[16] 平新乔、杨慕云：《消费信贷违约影响因素的实证研究》[J]，《财贸经济》2009 年第 7 期。

[17] 孙燕：《随机效应 Logit 计量模型的自适应 Lasso 变量选择方法研究》[J]，《数量经济技术经济研究》2012 年第 12 期。

[18] 王来福、郭峰：《中国住房抵押贷款信用风险：理论分析与实证研究》[J]，《数学的实践与认识》2009 年第 12 期。

[19] 王大荣、张忠占：《线性回归模型中变量选择方法综述》[J]，《数理统计与管理》2012 年第 4 期。

[20] 杨显爵、林左裕、陈震远、陈震武、陈宗豪：《小额信贷之违约概率模型：特别考虑异质性》[J]，《浙江大学学报（人文社会科学版）》2008 年第 3 期。

（责任编辑：陈星星）