Advances in Applied Mathematics 应用数学进展, 2023, 12(2), 563-573
Published Online February 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.122059

Hans汉斯

基于Multiple-Lasso-Logistic回归模型的车险索赔概率预测

马雨星

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年1月16日；录用日期：2023年2月11日；发布日期：2023年2月22日

摘　要

近年来，Logistic回归模型在非寿险精算科学中得到了广泛的应用，本文对法国的一组车险索赔数据，采用Lasso及其扩展方法结合Logistic模型建立车险索赔预测模型，同时引入了惩罚权重，并与Lasso-logistic回归模型和Logistic回归模型进行比较，结果表明：模型综合性能最优的是multiple-Lasso-Logistic回归模型。并在此基础上筛选出了预测性能最强的因子水平，同时对响应变量作用相同的水平进行了融合，有效地降低了变量维度。

关键词

Logistic回归模型，Lasso及其扩展方法，索赔概率，惩罚权重

Prediction of Automobile Insurance Claim Probability Based on Multiple-Lasso-Logistic Regression Model

Yuxing Ma

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Jan. 16th, 2023; accepted: Feb. 11th, 2023; published: Feb. 22nd, 2023

Abstract

In recent years, Logistic regression model has been widely used in non-life insurance actuarial science. This paper uses Lasso and its extension method combined with Logistic regression model to establish a prediction model for vehicle insurance claims based on a group of vehicle insurance

文章引用: 马雨星. 基于 Multiple-Lasso-Logistic 回归模型的车险索赔概率预测[J]. 应用数学进展, 2023, 12(2): 563-573.
DOI: 10.12677/aam.2023.122059

claim data in France, and introduces penalty weight. Compared with Lasso-logistic regression model and Logistic regression model, the results show that the multiple-Lasso-Logistic regression model has the best comprehensive performance. On this basis, the factor level with the strongest prediction performance is selected, and the level with the same effect of response variables is fused, which effectively reduces the variable dimension.

Keywords

Logistic Regression Model, Lasso and Its Extension Method, Claim Probability, Penalty Weight

1. 引言

精算学是一个在不断发展的领域，涉及保险风险的评估，其中车辆保险是为汽车、卡车、摩托车和其他公路车辆设计的保险，它不仅为交通碰撞造成的车辆损坏和人身伤害提供财务保护，而且还对冲了交通事故中可能出现的责任车辆保险。随着科技的进步，汽车在人民的生活中已越来越普及了，保费也在持续增长，车辆保险在非寿险业务中发挥着极其重要的作用，因此研究和改善汽车保险问题已成为学术界共同关注的问题，保险公司对投保人的纯保费由两部分组成：索赔概率和期望索赔额，本文对车险索赔概率进行研究。建立预测模型时，很少有文献会考虑多重共线性，模型的过拟合，即使有研究结合了 Lasso 和 Logistic 模型来消除多重共线性与模型过拟合，却也忽略了预测因子间水平结构的不同，对所有类型的预测因子都施加相同的惩罚，我们的研究考虑到预测因子具有不同的水平结构，对不同类型的预测因子选择最适合的正则设置，同时引入了惩罚权重，来提高惩罚的性能。

在现有文献中，Logistic 回归模型在保险与精算领域的应用受到了越来越多的关注，近几年已经取得了一些较好的研究成果，张连增和孙维伟(2012) [1]年运用 Logistic 模型分析了车险索赔概率的影响因素，通过分析得出驾驶人年龄、行驶区域、车辆类型、车辆价值等因素对于汽车是否出险具有显著影响。孟生旺等(2017) [2]对车险索赔概率分别建立了神经网络模型与广义线性模型，对以上两种模型进行比较，得出结论：神经网络模型对数据要求小，预测精度高，但缺点是建模类似一个黑匣子，结果缺乏可解释性。广义线性模型操作简单，可解释性强，但模型对原始数据有较多要求。王冲(2017) [3]分别从汽车因素、驾驶人因素两大方面进行研究，基于汽车是否出险，建立了 Logistic 回归模型，探究影响汽车出险的因素以及各变量的重要程度，并进一步运用树模型分析了各因素之间的交互效应，从而对 Logistic 回归模型进行了优化。郑俊卿(2018) [4]将 Lasso 回归和 GLM 模型相结合来预测车险索赔频率，将改进后的模型与 GLM 模型进行比较，结论显示改进后的模型的性能更好。卢志义和蔡静(2017) [5]将驾驶员的性别、车型等 8 个变量作为预测因子，分别建立了车险索赔概率估计的 Logistic 回归模型和广义可加模型，结果表明，对于离散型费率因子占绝大多数的车险数据，广义可加模型并不具有明显的优势，因此在车险费率厘定实务中，若离散型费率因子较多，应选择结构相对简单的 Logistic 回归模型。Oelker 和 Tutz (2017) [6]将多类型正则化扩展到 GLM 模型中，使得惩罚迭代加权最小二乘(PIRLS)算法适用于正则化 GLM 模型。Devriendt 等(2021) [7]设计了一种更有效的校准策略，适用于不同预测因子类型和更一般的损失函数的正则化，将方法应用在车险索赔次数的预测中，其效果要优于广义可加模型。很多研究在建立车险索赔概率预测模型时，会忽略变量之间的多重共线性和模型的过拟合，或者使用 Logistic 模型结

合单一类型的 Lasso 回归来解决变量的多重共线性或模型过拟合，忽略了变量之间水平结构的不同。同时 Lasso 回归只是孤立的考虑各个变量，未考虑到各个变量之间的有序性，因此，本文使用 Lasso 及其扩展方法结合 Logistic 回归模型建立索赔概率预测模型，并将其与 Lasso-Logistic 回归模型与 Logistic 回归模型进行比较。

2. 模型介绍

2.1. Logistic 回归模型

Logistic 回归模型是一种因变量为分类变量的模型。在车险保单组合中，假设个体保单之间相互独立，每份保单是否发生索赔可以表示为一个二分类随机变量，服从伯努利分布，发生索赔的概率函数如下式所示。

$$P(Y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\cdots+\beta_Jx_J)}}, \tag{1}$$

将上式进一步变换可得

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1x_1+\cdots+\beta_Jx_J, \tag{2}$$

其中，$x=(x_1,x_2,\cdots,x_J)$ 为车险索赔概率的影响因素，J 表示影响因素的个数，在本文中共有 9 个影响因素，因此 $J=9$，其中 x_1 表示车龄，x_2 表示驾驶人年龄，x_3 表示奖惩水平，x_4 表示人口密度，x_5 表示车辆马力，x_6 表示车辆品牌，x_7 表示燃料类型，x_8 表示居住地区，x_9 表示暴露期，P 表示索赔发生的概率，β_0 表示截距项，β_j 表示第 j 个车险索赔概率影响因素的回归系数，Y 为响应变量，$Y=0$ 表示发生索赔，$Y=1$ 表示未发生索赔。

2.2. Lasso 及其扩展

Lasso：Lasso 惩罚将 L_1 范数应用于预测系数，其表达式如下

$$g_{Lasso}(\beta_j)=\sum_{i=1}^{p_j}w_{ji}|\beta_{ji}|, \tag{3}$$

β_{ji} 表示第 $j(j=1,2,\cdots,J)$ 个车险索赔概率影响因素的第 i 个水平，p_j 表示第 $j(j=1,2,\cdots,J)$ 个车险索赔概率影响因素中的水平个数，w_{ji} 表示第 $j(j=1,2,\cdots,J)$ 个车险索赔概率影响因素的第 i 个水平的惩罚权重，对于由一个系数表示的连续预测变量或二元预测变量，最重要的预测因子获得非零系数。

Grouplasso：Grouplasso 惩罚使用 L_2 范数，其表达式如下

$$g_{grplasso}(\beta_j)=w_j\sqrt{\sum_{i=1}^{p_j}\beta_{ji}^2}, \tag{4}$$

该惩罚适用于确定 β_j 是否具有足够的整体预测能力，Grouplasso 惩罚对于选择有序或名义预测变量特别有用，当将其应用于有序或名义预测值时，Grouplasso 不需要参考类别。

Fuse lasso：将预测值中的连续级别进行分组，Fuselasso 惩罚会对后续系数之间的差异进行 L_1 惩罚，其表达式如下

$$g_{flasso}(\beta_j)=\sum_{i=2}^{p_j}w_{j(i-1)}|\beta_{ji}-\beta_{j(i-1)}|, \tag{5}$$

这种惩罚适用于有序变量和编码为有序变量的连续型预测变量，以捕捉非线性效应，对于有序变量，相邻预测变量的作用往往相同或者相近，因此 Fuselasso 在进行变量选择的同时，对相邻参数的差进行收

缩，以达到相邻参数相同或相近的目的。

Generalized fused lasso：Generalized fused lasso 允许用户设置一个图 M (在 R 语言中的 smurf 包中自动实现)，该图指示哪些系数差应被正则化，其表达式如下

$$g_{gflasso}\left(\beta_j\right)=\sum_{(i,l)\in M} w_{j(il)}\left|\beta_{ji}-\beta_{jl}\right|, \tag{6}$$

对于空间预测变量(二维预测变量)，Generalized fused lasso 调整了共享物理边界的城市的系数差异，对于没有任何潜在结构的名义预测值，可以使用图对所有可能的系数差异进行正则化。

2.3. 自适应惩罚权重

一个好的变量选择方法应该具有神域性质，该性质包含两层含义：模型估计的相合性，参数估计的相合性。然而 Lasso 并不具备这些性质，为了克服 Lasso 这一性质，Zou (2006) [8]提出了自适应 Lasso，自适应 Lasso 被定义为：$w_{j,i}^{ad}=\left|\hat{\beta}_{j,i}\right|^{-\gamma}$，其中 $\gamma>0$ 为调谐参数。直观上，这些权重以初始估计量的形式将惩罚“适应”于数据驱动的先验信息。在本文中，我们参照 Gertheiss 和 Tutz (2010) [9]那样设置 $\gamma=1$，并在表 1 中列出特定惩罚的自适应权重。

Table 1. Table of penalty weights
表 1. 惩罚权重表

Penalty name	Lasso	Group lasso	Fused lasso	Generized fused lasso
$w_j^{(ad)}$	$w_{ji}^{ad}=\left\|\hat{\beta}_{ji}\right\|^{-1}$	$w_{ji}^{ad}=\left\|\hat{\beta}_{ji}\right\|_2^{-1}$	$w_{j(i-1)}^{ad}=\left\|\hat{\beta}_{ji}-\hat{\beta}_{j(i-1)}\right\|^{-1}$	$w_{j(il)}^{ad}=\left\|\hat{\beta}_{ji}-\hat{\beta}_{jl}\right\|^{-1}$

3. 实证分析

3.1. 变量来源及说明

本文采用的是法国的一组车险索赔数据，来源于 R 中的 maidrr 包，该数据集包含 668,892 个观测值，10 个变量：nclaims 表示索赔频率，power 表示车辆马力(马力越大，极速越大)，bm 表示奖惩水平(奖惩水平越低表明驾驶员的索赔历史记录越好)，brand 表示车辆品牌，fuel 表示燃料类型，agep 表示驾驶人年龄，agec 表示车龄，expo 表示风险暴露期，Popden 表示人口密度，region 表示居住地区，表 2 为变量说明表：

Table 2. Variable description table
表 2. 变量说明表

变量	符号	类型	说明
x_1	agec	有序变量	车龄：0~100
x_2	ageph	有序变量	驾驶人年龄：18~100
x_3	bm	有序变量	奖惩水平：50~230
x_4	popden	有序变量	居住城市的人口密度：1~27,000
x_5	Power	有序变量	车辆马力：12 个有序级别
x_6	brand	名义变量	车辆品牌：11 个水平
x_7	fuel	二元变量	燃料类型：柴油或汽油
x_8	region	名义变量	居住地区：22 个水平
x_9	expo	有序变量	暴露期：0~1
y	I	二元变量	是否发生索赔：取值为 0 或 1

3.2. 因素分析

对于车险索赔概率的因素分析，连续型和有序型变量运用箱线图进行初步分析。

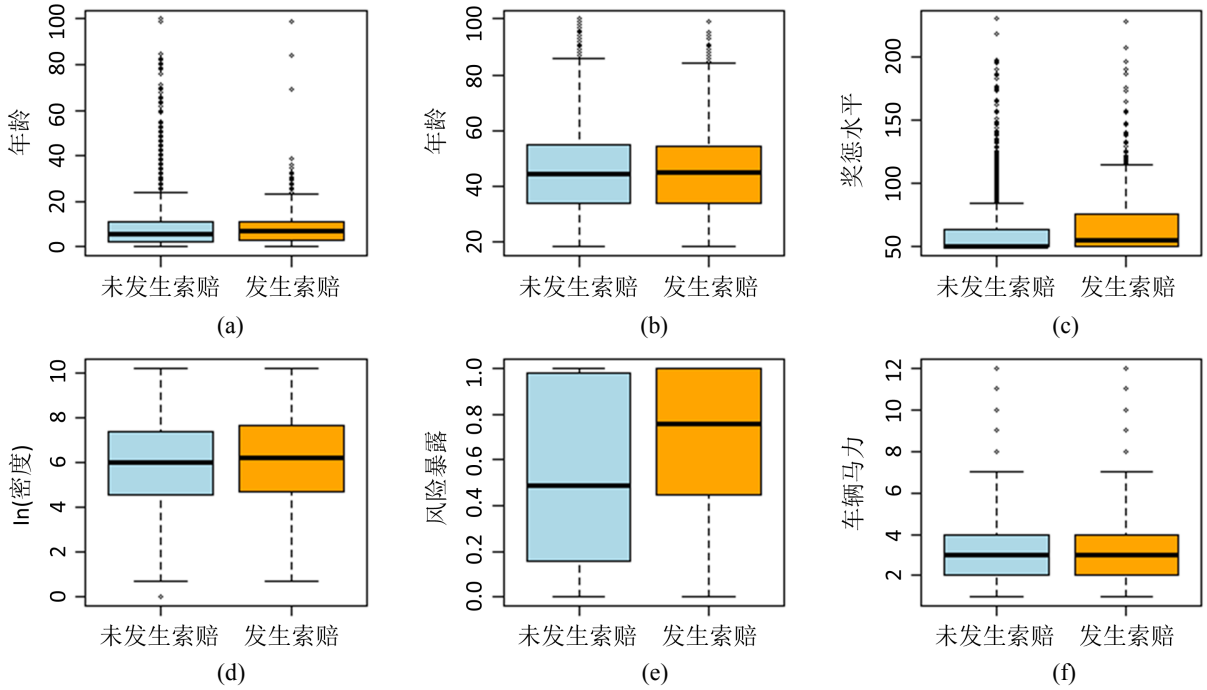

Figure 1. Automobile insurance claim factor diagram
图 1. 车险索赔因素图

图 1 中箱子中间的一条线，是数据的中位数，代表了样本数据的平均水平，箱子的上下限，分别是数据的上四分位数和下四分位数，因此，箱子的宽度在一定程度上反映了数据的离散程度，根据图 1(a)、图 1(b)、图 1(f)可以看出未发生索赔的驾驶人驾龄、年龄、车辆马力的中位数和发生索赔的驾驶人驾龄、年龄、车辆马力的中位数大致相等，且箱体宽度相同，因此驾驶人驾龄、年龄、车辆马力在索赔和未发生索赔两种情况下的分布大致相同，根据图 1(c)可以看出发生索赔的驾驶人奖惩水平的中位数要高于未发生索赔的驾驶人奖惩水平的中位数，且在发生索赔情况下的箱体要宽于未发生索赔情况下的箱体，因此在发生索赔情况下奖惩水平的分布更加离散，根据图 1(d)可以看出发生索赔的驾驶人居住地区密度的整体水平要略高于未发生索赔的驾驶人居住地区密度的整体水平，因此人口密度高的地区的驾驶人更容易发生索赔，根据图 1(e)可以看出发生索赔驾驶人的风险暴露期要明显高于未发生索赔驾驶人的风险暴露期，值得注意的是发生索赔和未发生索赔二者的样本分布是不平衡的，因此以上分析的差异是否显著需要通过进一步建模进行判断。

3.3. 车险索赔概率预测模型的建立

Logistic 回归的损失函数：

$$L(\beta) = -\frac{1}{n}\sum_{t=1}^{n}\left(y_t x_t \beta - \log\left(1+\exp\left(x_t\beta\right)\right)\right), \tag{7}$$

n 表示保单数量，y_t 表示第 $t(t=1,2,\cdots,n)$ 份保单是否发生索赔，$\beta=\left(\beta_0,\beta_1,\cdots,\beta_j\right)$ 表示参数向量，$x_t\left(t=1,2,\cdots,n\right)$ 表示第 t 份保单的自变量观测值。

Multiple-Lasso-Logistic 即在 Logistic 回归的损失函数上添加惩罚函数，根据不同的变量类型，在式(3)、(4)、(5)、(6)中选择适合的惩罚函数。在进行参数估计时，通过压缩系数实现对变量的选择，将没有预测能力的变量压缩为 0，得到解释能力较强的变量，则 Multiple-Lasso-Logistic 回归模型中的系数估计值 $\hat{\beta}$ 可以写成如式：

$$\hat{\beta}=\arg\min\left[L(\beta)+\lambda\left(\sum_{j\in\{7\}}g_{grplasso}(\beta_j)+\sum_{j\in\{1,2,3,4,5,9\}}g_{flasso}(\beta_j)+\sum_{j\in\{6,8\}}g_{gflasso}(\beta_j)\right)\right], \tag{8}$$

根据各个变量的水平结构，我们对二元预测变量燃料类型(当 $j=7$ 时)应用 Group lasso 惩罚，对连续变量车龄，驾驶人年龄，车辆马力，奖惩水平，人口密度，暴露期(当 $j=(1,2,3,4,5,9)$ 时)进行分箱处理后应用 Fused lasso 惩罚，其考虑到了变量之间的次序作用，对名义变量居住地区、车辆品牌(当 $j=(6,8)$ 时)应用 Generalized fused lasso，同时引入惩罚权重，本文使用表 1 中的自适应惩罚权重。

Fused lasso 适用于有序型变量或编码为有序变量的连续型预测变量，因此本文需要对连续型变量进箱处理，进行分箱处理时须确保每一水平中的样本量不可以过少，若某一水平的样本量过少则会影响预测的精度，同时有效的分箱处理也可以提高模型的预测效果，若人为的采取等距分箱，或是主观的进行分箱，得到的不一定是最优分箱，因此本文采用 R 中的 smbining 包，基于监督离散化，利用递归分区来将数字特征分类，根据 smbinning 自带的 ctree 算法进行分裂，找出最优分割点，然后计算 IV 值，计算方式如下

$$\mathrm{IV}=\sum_{k=1}^{N}\mathrm{IV}_k\ , \tag{9}$$

$$\mathrm{IV}_k=\left(\frac{\mathrm{Positive}_k}{\mathrm{Positive}}-\frac{\mathrm{Negitive}_k}{\mathrm{Negitive}}\right)\ln\left(\frac{\dfrac{\mathrm{Positive}_k}{\mathrm{Positive}}}{\dfrac{\mathrm{Negative}_k}{\mathrm{Negative}}}\right), \tag{10}$$

其中 $\mathrm{Positive}_k$ 表示经过 smbining 包分箱后第 k 组中发生索赔的样本数量，Positive 表示所有组中发生索赔的样本数量，$\mathrm{Negative}_k$ 表示第 k 组中未发生索赔的样本数量，Negative 表示所有组中未发生索赔的样本数量。N 表示分组的数量，经过计算可知，车龄的 IV 值为 0.0446，奖惩水平的 IV 值为 0.1190，年龄的值 IV 值为 0.0198，人口密度的 IV 值为 0.0131，可知车龄的分箱效果最好，人口密度的分箱效果最差，分箱情况如表 3。

Table 3. Bin table for continuous variables
表 3. 连续变量的分箱表

	分组频数							
x_1 频数	≤ 0 (53,924)	≤ 1 (70,713)	≤ 3 (108,458)	≤ 4 (43,031)	≤ 12 (266,187)	≤ 14 (51,632)	≤ 16 (38,679)	> 16 (36,268)
x_2 频数	≤25 (38,493)	≤ 40 (233,382)	> 40 (397,017)					
x_3 频数	≤ 51 (393,811)	≤ 57 (51,936)	≤ 63 (37,747)	≤ 68 (42,306)	≤ 90 (98,784)	> 90 (44,308)		
x_4 频数	≤ 40 (80,784)	≤ 393 (254,573)	≤ 526 (34,125)	≤ 730 (34,786)	> 730 (264,624)			

在对连续型变量分箱后，采用 10 折交叉验证法来确定调节参数 λ，数据集被分成 10 个不相交的集合，我们将其中 9 个集合组成的训练样本来建立模型，然后使用剩余的集合(即验证样本)来计算误差度量。

重复 10 次，这样每个集合正好用作验证样本一次，这 10 个误差测量值的平均值用作 λ 值的误差测量值。

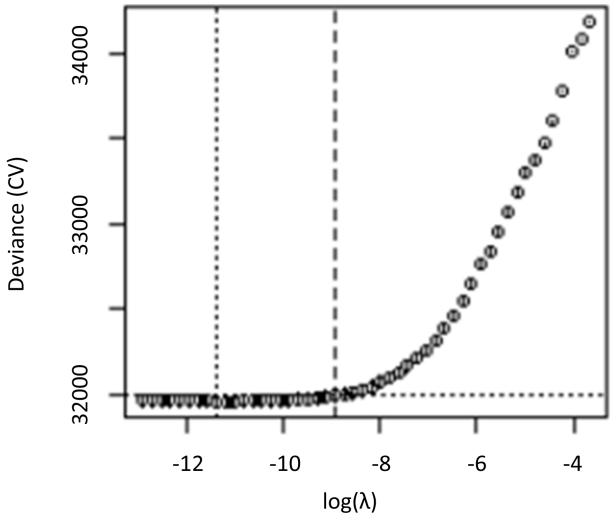

Figure 2. Diagram of harmonic parameters and deviations
图 2. 调和参数和偏差的关系图

图 2 纵轴表示模型的偏差，横轴为 $\log(\lambda)$，图中每个点表示调和参数 λ 取值对应的偏差值，随着 λ 的变大，惩罚力度也会变大，于是便有更多的系数值被压缩为 0，因此需要找到一个合适的 λ 值，可以看出随着调和参数 λ 值的增加，偏差逐渐变小，图中两条虚线分别表示 lambda.min 和 lambda.1se，本文选取后者，即一个方差范围内最精简模型的 λ 值，即 $\lambda = 0.0001325699$，得到 λ 值后，利用 R 软件的 smurf 包建立多类型 Lasso 的 Logistic 回归模型，模型的参数估计值如表 4 所示。

从表 4 可以看出，有参数估计和参数再估计两列数据，参数估计是有惩罚的情况下得到的参数估计

Table 4. Parameter estimation table
表 4. 参数估计表

Parameter	Estimated	Re-estimated	Parameter	Estimated	Re-estimated
Intercept	−4.18950	−4.17241	x_{68}	0.11488	0.02728
x_{11}	0.00930	0.08820	x_{69}	0.10249	0.02357
x_{12}	0.01724	0.15796	x_{610}	0.17348	0.09104
x_{13}	0.01724	0.15796	x_{611}	0.12034	0.04494
x_{14}	0.01724	0.15796	x_{71}	0.03634	0
x_{15}	−0.15893	−0.03381	x_{72}	−0.03634	−0.12636
x_{16}	−0.27250	−0.18409	x_{81}	0	0
x_{17}	−0.27250	−0.18409	x_{82}	0	0
x_{21}	−0.00471	−0.13040	x_{83}	0	0
x_{22}	0.38298	0.29767	x_{84}	0	0
x_{31}	0.38235	0.41933	x_{85}	0	0
x_{32}	1.06089	1.08860	x_{86}	0	0
x_{33}	0.56812	0.53534	x_{87}	0	0
x_{34}	0.84035	0.84767	x_{88}	0	0

Continued

x_{35}	1.47813	1.47532	x_{89}	0	0
x_{41}	0.06381	0.13264	x_{810}	0	0
x_{42}	0.15616	0.23155	x_{811}	0	0
x_{43}	0.15616	0.23155	x_{812}	0	0
x_{44}	0.23625	0.33717	x_{813}	0	0
x_{51}	0.06584	0.10830	x_{814}	0	0
x_{52}	0.06584	0.10830	x_{815}	−0.13936	−0.29419
x_{53}	0.06584	0.10830	x_{816}	0.10391	0.24914
x_{54}	0.06584	0.10830	x_{817}	0.17795	0.20794
x_{55}	0.17222	0.26283	x_{818}	0	0
x_{56}	0.17222	0.26283	x_{819}	0	0
x_{57}	0.17222	0.26283	x_{20}	0	0
x_{58}	0.17222	0.26283	x_{21}	0	0
x_{59}	0.17222	0.26283	x_{91}	−1.49205	−1.54313
x_{59}	0.17222	0.26283	x_{92}	−0.65263	−0.73935
x_{510}	0.17222	0.26283	x_{93}	−0.31357	−0.45257
x_{61}	0.07747	0	x_{94}	−0.23871	−0.38460
x_{62}	0.10147	−0.01309	x_{95}	0	0
x_{63}	0.23292	0.13926	x_{96}	0	0
x_{64}	−0.33466	−0.41231	x_{97}	0.37788	0.37030
x_{65}	0.11864	0.02594	x_{98}	0.61880	0.60131
x_{66}	−0.20876	−0.29601	x_{99}	0.53604	0.51005
x_{67}	0.07558	−0.00398			

值，因此会有一定偏差，再估计值根据参数估计的结果，将系数被压缩为 0 的变量去除，然后不设置惩罚，重新估计的系数将具有与正则化估计相同的非零和融合系数，但不会有偏差。

根据上表可得：

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\sum_{k=1}^{7}\beta_{1k}x_{1k}+\sum_{k=1}^{2}\beta_{2k}x_{2k}+\sum_{k=1}^{5}\beta_{3k}x_{3k}+\sum_{k=1}^{4}\beta_{4k}x_{4k}+\sum_{k=1}^{11}\beta_{5k}x_{5k}+\sum_{k=1}^{11}\beta_{6k}x_{6k}+\sum_{k=1}^{2}\beta_{7k}x_{7k}+\sum_{k=1}^{21}\beta_{8k}x_{8k}+\sum_{k=1}^{9}\beta_{9k}x_{9k},\tag{11}$$

将上表参数可视化，可以更加清楚地看出各个变量对索赔是否发生的影响。

根据图 3(a)可以看出，相对于驾龄低的驾驶人，驾龄高高的驾驶人发生索赔概率更低，图 3(b)显示了年老的司机发生索赔的概率更高，因此保险公司需承担更高的风险。图 3(c)的趋势总体是呈上升的，具有高奖惩水平(索赔历史记录越差)的驾驶人发生索赔的概率越高，图 3(d)表明人口密度高的地方更容易发生索赔，因此对保险公司来说风险更高，图 3(e)显示动力强劲的汽车比动力较弱的汽车具有更高的风险，图 3(h)显示居住地区的系数大多数被压缩为了 0，因此对索赔是否发生的影响不大，但可以看出编号为 R73 的地区不容易发生索赔，编号为 R74 和 R82 的地区相对于其他区更容易发生索赔，从参数估计图中可以看出很多水平进行了融合，与标准 GLM 相比可以很好地(并且自动地)降低了维度(图 3)。

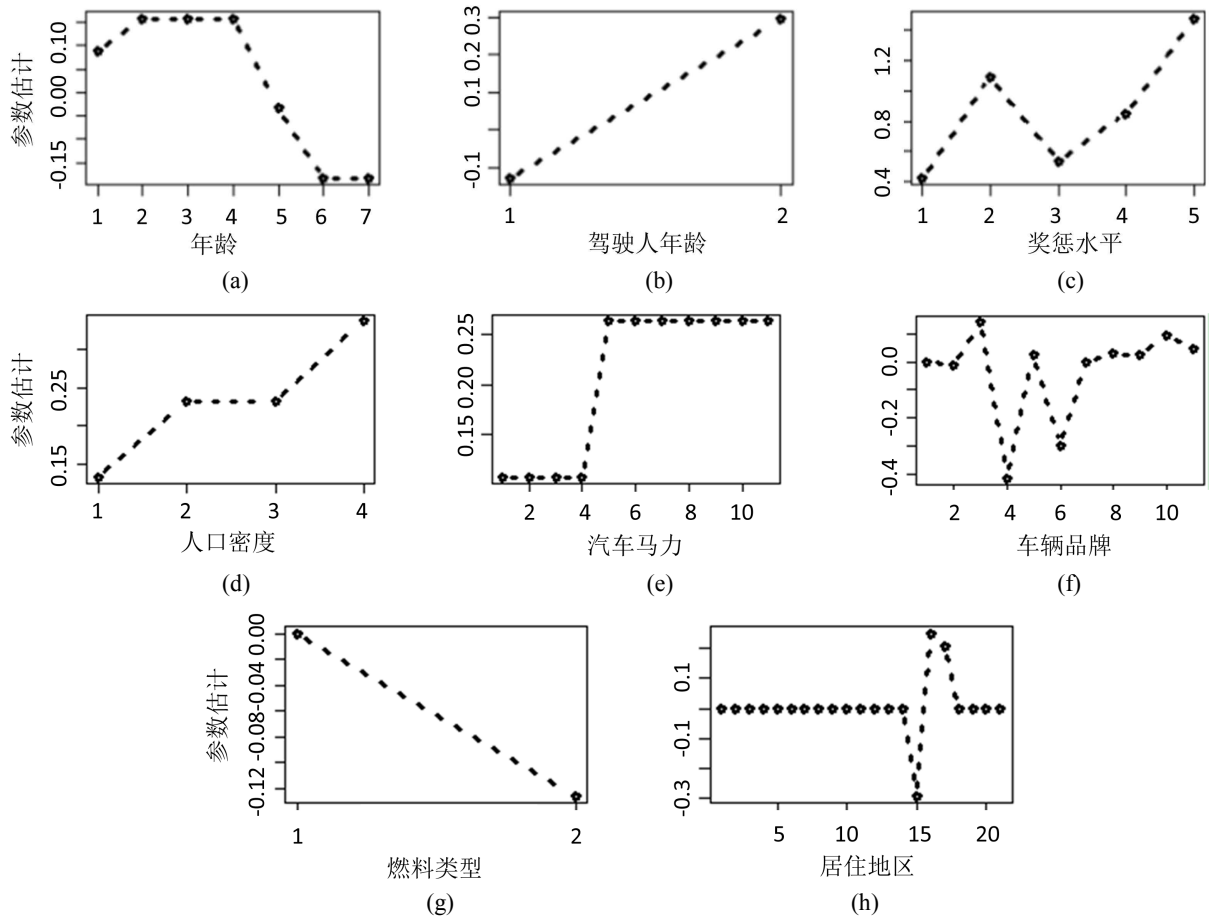

Figure 3. Parameter estimation plot
图 3. 参数估计图

3.4. 模型的评价与比较

本文在比较不同模型的拟合效果时主要采用赤池信息准则(AIC)，和贝叶斯信息准则(BIC)，即如下：

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{12}$$

$$\mathrm{BIC} = k\ln(n) - 2\ln(L), \tag{13}$$

其中 k 是模型参数个数，L 是对数似然函数，n 为样本数量，赤池信息准则(AIC)提供了权衡估计模型复杂度和拟合数据优良性的标准，BIC 信息准则与 AIC 相似，用于模型选择，AIC 和 BIC 均引入了与模型参数个数相关的惩罚项，BIC 的惩罚项比 AIC 的大，考虑了样本数量，表 5 为各个模型的 AIC，BIC 值，可以看出 Multiple-Lasso-Logistic 模型表现的最好。

Table 5. AIC and BIC comparison table for each model
表 5. 各个模型的 AIC 和 BIC 比较表

模型	AIC	BIC
Lasso-Logistic	150,736	150,992
Multiple-Lasso-Logistic	149,618	150,052
Logistic	149,628	150,407

AUC 是二分类模型的评价指标，是 ROC 曲线下方的面积，ROC 曲线由两个变量 TPR 和 FPR 组成，这个组合以 FPR 对 TPR，即是以代价对收益。

x 轴为假阳性率(FPR)：所有的负样本中，分类器预测错误的比例

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{14}$$

y 轴为真阳性率(TPR)：所有的正样本中，分类器预测正确的比例

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{15}$$

AUC 的范围为 0 到 1 之间，AUC 值越接近 1，模型的预测效果越好，表 6 表示各个模型的 AUC。

Table 6. AUC comparison table for each model
表 6. 各个模型的 AUC 比较表

模型	Lasso-Logistic	Multiple-Lasso-Logistic	Logistic
AUC	0.6879	0.7035	0.7035

标准的 Logistic 回归模型会将所有变量纳入模型中，因此也会纳入一些不显著的变量，可能会扭曲参数的估计结果，Multiple-Lasso-Logistic 会针对到变量的水平结构，对其应用合适的正则化方法，将预测能力弱的变量压缩为 0，筛选出预测能力最强的变量，同时还能将作用相同水平进行融合，有效地降低了变量维度，根据表 5，表 6 可以看出，Multiple-Lasso-Logistic 回归模型的 AIC 和 BIC 都是最小的，其 AUC 值和 Logistic 回归模型的大小相同，结合以上三个模型评价指标可得 Multiple-Lasso-Logistic 回归模型的性能最优。

4. 小结

研究车险索赔概率时，现有文献中，对于处理多重共线性以及模型过拟合问题时，大多文献未考虑到车险因子水平结构的不同，本文针对不同结构的车险因子水平，采用 Lasso 及其扩展方法选择具有预测能力的因子。本文对法国的一组车险索赔数据分别建立了 logistic 回归模型，Lasso-Logistic 回归模型，Multiple-Lasso-Logistic 回归模型，结果表明 Multiple-Lasso-Logistic 回归模型性能最优。标准的 logistic 回归模型会将不显著的变量也纳入模型中，往往会扭曲参数估计的结果，因此本文使用 lasso 及其扩展方法来改进 logistic 回归模型，相比于 lasso 回归考虑到了变量之间水平结构的不同。

本文利用 AIC，BIC，AUC 三个评价指标对模型的性能进行评价，Multiple-Lasso-Logistic 回归模型不仅可以筛选出预测性能最高的变量，且可以将作用相同的水平进行融合，有效地降低了变量维度，使模型变得更精炼。同时 Multiple-Lasso-Logistic 回归模型可以帮助保险公司有效识别风险，进行风险分类，对不同风险级别的被保险人采用不同的保费策略提供一定的参考价值。

参考文献

[1] 张连增, 孙维伟. 车险索赔概率影响因素的 Logistic 模型分析[J]. 保险研究, 2012(7): 16-25.

[2] 孟生旺, 李天博, 高光远. 基于机器学习算法的车险索赔概率与累积赔款预测[J]. 保险研究, 2017(10): 42-53.

[3] 王冲. 汽车出险影响因素分析[D]: [硕士学位论文]. 北京: 北京理工大学, 2017.

[4] 郑俊卿. 基于 Lasso 回归的 GLM 模型及其在车险费率厘定中的应用研究[D]: [硕士学位论文]. 青岛: 山东科技大学, 2018.

[5] 卢志义, 蔡静. 车险费率厘定的索赔概率预测模型及其比较分析[J]. 河北工业大学学报, 2017, 46(3): 56-62.

[6] Oelker, M.R. and Tutz, G.A. (2017) Uniform Framework for the Combination of Penalties in Generalized Structured Models. *Advances in Data Analysis and Classification*, **11**, 97-120. https://doi.org/10.1007/s11634-015-0205-y

[7] Devriendt, S., Katrien, A., Tom, R. and Roel, V. (2021) Sparse Regression with Multi-Type Regularized Feature Modeling. *Insurance Mathematics and Economics*, **96**, 248-261. https://doi.org/10.1016/j.insmatheco.2020.11.010

[8] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[9] Gertheiss, J. and Tutz, G. (2010) Sparse Modeling of Categorial Explanatory Variables. *The Annals of Applied Statistics*, **4**, 2150-2180. https://doi.org/10.1214/10-AOAS355