

均值计数模型下汽车保险索赔频数的估计方法

赵晓兵, 刘 伟

(浙江财经大学数学与统计学院, 浙江 杭州 310018)

摘 要: 汽车保险的索赔频数预测问题是非寿险精算理论和应用研究的一个重要内容。但是, 在含有高维附加信息的情形下, 传统的估计方法就不再适用。本文在均值计数模型基础上, 利用凸惩罚函数进行变量选择, 找到影响车险索赔频数的显著性因子, 并通过模拟和实例分析来评价该模型和所提出的方法的可行性。

关键词: 汽车保险; 均值计数模型; 凸惩罚; 变量选择; 估计方程

中图分类号: F840.6 文献标识码: A 文章编号: 1004-4892(2015)02-0044-06

一、引 言

汽车商业保险是对机动车辆由于自然灾害或意外事故所造成的人身伤亡或财产损失承担赔偿责任的一种保险业务。随着汽车数量的猛增, 车险市场呈现出快速发展的态势。汽车保险更是财产保险的第一大险种, 部分公司的汽车保险保费收入占其财产保险总保费收入的60%以上。关于汽车保险定价方法的研究一直以来都是非寿险精算理论及应用研究的重点内容。

在目前的汽车保险定价实务中, 对车险索赔频率和索赔强度的预测是两个主要研究问题, 流行的研究方法是利用广义线性模型方法^{[1][2]}。虽然广义线性模型有现成的统计软件可用, 也可以对参数估计的结果进行直观的解释, 但是, 该方法需要假定已知因变量和解释变量之间的某种联系函数, 而目前采用的函数形式却比较有限。随着现代统计方法的大量出现, 以及数据收集方式的更新, 使得新类型的数据往往包含大范围的附加信息, 即所谓的“高维协变量”^[3]。在这种背景下, 传统的广义线性模型往往不再适用。而且由于广义线性模型不能自动识别解释变量之间的交互作用, 导致建模过程比较耗时。除广义线性模型之外, 神经网络模型也是研究汽车保险索赔问题的常用研究方法之一。但神经网络模型的计算较为复杂, 同时也很难对协变量的回归系数给出直观的解释(Faraway, 2006; Werner and Modlin, 2010; 孟生旺, 2007)^{[4][5][6]}。

针对现有汽车保险索赔频数估计方法中存在的局限, 本文基于澳大利亚 MAA 公司(The Motor Accidents Authority)的一组综合险(comprehensive insurance)索赔数据, 将 Wang、Qin and Chiang (2001)^[7]以及 Huang and Wang(2004)^[8]的模型推广到允许含有高维协变量存在的情形, 在此基础上提出一个新的评估方法。该模型有两个显著特点: 一是允许高维协变量的存在, 可以通过变量选择得到模型的稀疏表达, 找到影响索赔频数的显著性因子, 提高模型整体的预测精度。二是对未知

收稿日期: 2014-08-27

基金项目: 国家自然科学基金资助项目(11271317); 浙江省自然科学基金资助项目(LY14A010022); 浙江省哲学与社会科学规划资助项目(12JCJJ17YB)

作者简介: 赵晓兵(1968-), 男, 四川平昌人, 浙江财经大学数学与统计学院教授; 刘伟(1987-), 男, 山东泰安人, 浙江财经大学数学与统计学院硕士生。

的基准函数不进行任何参数假定, 并且在降维的过程中不需要知道基准函数的具体形式, 以便对车险索赔频数做出更稳健的估计。

二、模 型

在索赔频数或者复发事件研究中, 我们常常采用 Cox 型强度函数的计数过程。假定因变量 $N_i (i = 1, 2, \dots, n)$ 为汽车保险索赔频数, 解释变量 $X_{i1}, X_{i2}, \dots, X_{ip}$ 为影响车险索赔频数的风险因子。为了分析该索赔数据, Huang and Wang (2004)^[8]提出了如下模型:

$$\lambda(t | X_i) = \lambda_0(t) \exp(X_i^T \beta) \quad (i = 1, 2, \dots, n) \quad (1)$$

其中, $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$, $\beta_i = (\beta_1, \beta_2, \dots, \beta_p)^T$, $t = Y_i \wedge \tau$ 为观测时间, Y_i 为删失时间, τ 为观测的终止时间, $\lambda_0(t)$ 是未知的基准函数(baseline), $\lambda(t | X_i)$ 是强度率函数。上述模型常常被称为 Cox 型比例危险模型。

然而此模型也存在一些局限, 例如, 我们常常需要假定其协变量是低维的。当含有高维协变量时, 该模型往往不再适用。基于此, 本文对该模型进行一般化推广, 即允许有高维协变量的存在。Zhao and Zhou(2014)^[9]对含有高维协变量的 Cox 模型下的系数估计方法进行了深入研究, 提出如下的多指标模型:

$$\lambda(t | X_i) = \lambda_0(t) \exp(\psi(X_i^T \beta)) \quad (i = 1, 2, \dots, n) \quad (2)$$

其中, Ψ 为完全未知的联系函数。首先利用非参数方法对未知的基准函数 $\Lambda_0(t)$ 做出估计, 其次使用充分降维(sufficient dimension reduction-SDR)获得协变量的中心降维子空间的结构维数和基方向, 最后通过局部回归估计完全未知的联系函数 Ψ 。

注意到 Zhao and Zhou(2014)^[9]对 $\Lambda_0(t)$ 的估计需要使用每次索赔发生的具体时间数据, 而在目前的精算实务中, 保险精算数据往往只含有累积的索赔次数, 而并不特别关心每次索赔具体发生的时间点。因此, 在本文中, 我们只需要对协变量进行降维, 而不再关注基准函数 $\Lambda_0(t)$ 的估计。假设一个均值计数模型, 即假设到时刻 t 为止的累积索赔次数 $N_i(t)$ 有如下的均值计数结构:

$$E(N_i(t) | X_i) = \Lambda_0(t) \exp(X_i^T \beta) \quad (i = 1, 2, \dots, n) \quad (3)$$

显然, (3) 式可以通过对(1)式两端对 t 积分得到, 其中 $\Lambda_0(t) = \int_0^t \lambda_0(u) du$, $N_i(t)$ 是第 i 个个体到 t 时刻的累积索赔频数。(1) 式中的 $\lambda(t | X)$ 即为模型(3) 中计数过程 $N_i(t)$ 的强度函数。

另外, SDR 可以有效克服高维协变量情形下“维数祸根”的影响, 且不需要对模型有任何参数假定, 在降维的过程中也充分考虑了响应变量的因素, 保留了更多的回归信息。但类似于主成分分析, SDR 是通过寻找自变量的若干线性组合来达到降维目的的, 因此我们不易得到降维系数的直观解释。为了找到影响汽车保险索赔频数的显著性因子, 赋予模型以直观的解释, 同时提高模型整体的预测精度, 本文考虑另一种方法, 即通过优化一个带“惩罚”函数的“损失”来达到变量选择的目的, 该方法也是目前文献中另外一个受到广泛重视的解决高维协变量问题的有效方法。受 Fan and Li(2001)^[10]惩罚对数似然函数思想的启发, 本文在模型(3)的基础上, 对 Sun and Wei(2000)^[11]提出的估计方程做出惩罚, 以得到 β 的稀疏估计。本文的显著优点在于: 一是可以允许有高维协变量的存在, 二是通过惩罚函数挑选显著性变量时不需要依赖基准函数 baseline。

注意到模型(1)和(3)虽有上述数学表达式上的联系, 但实际上它们却有很大的差别。模型(1)是一个基于非平稳泊松分布的计数过程, 模型(3)则为不需要关于分布作任何假设的均值计数模型。另外, 在估计方法上, 模型(1)和(2)均需要知道每次索赔发生的具体时间点, 而模型(3)却允许索赔发生的时间点完全未知。因此无论是在统计建模还是估计方法上, 模型(3)比模型(1)都更

具灵活性和更一般化。

三、模型估计

本节将利用凸惩罚函数方法来进行变量选择, 得到影响车险索赔频数的显著性因子及相应的系数估计。在模型(3)基础上, 为了得到参数向量 β 的估计, Sun and Wei(2000)^[11]提出了如下的无偏估计方程, 该方法的最大特点是不涉及未知的基准函数 $\Lambda_0(t)$, 从而不需要每次索赔的具体发生时间点。中心化协变量 X_i 后, 该估计方程定义如下:

$$W(\beta) = \sum_{i=1}^n N_i X_i \exp(-X_i^T \beta) \quad (i = 1, 2, \dots, n)$$

令其为 0, 即可得到 $\hat{\beta}$, Sun and Wei(2000)^[11]同时证明了该估计具有相合性与渐进正态性等优良性质。为了得到 β 的稀疏估计, Tong and He 等(2009)^[12]对该无偏估计加入惩罚, 得到如下形式的惩罚估计方程:

$$Q(\beta) = W(\beta) - n(p_{\lambda_1}'(|\beta_1|) \text{sgn}(\beta_1), \dots, p_{\lambda_d}'(|\beta_d|) \text{sgn}(\beta_d))'$$

其中, $p_{\lambda_j}'(|\beta_j|)$ 是给定的惩罚函数 p_{λ_j} 的一阶导数, 常用的惩罚函数包括 LASSO 惩罚、SCAD 惩罚、HARD 惩罚、岭回归、桥回归等形式。本文选择的惩罚函数为 LASSO 型惩罚, 即 $p_{\lambda_j}(|\beta_j|) = \lambda_j |\beta_j|$, λ_j 为调整参数, sgn 为符号函数, 令

$$A(\beta) = \frac{1}{n} \sum_{i=1}^n N_i \exp(-X_i^T \beta) X_i X_i'$$

$$\sum(\beta) = \text{diag}\{p_{\lambda_1}'(|\beta_1|)/|\beta_1|, \dots, p_{\lambda_d}'(|\beta_d|)/|\beta_d|\}$$

为了得到 β 的估计, Tong and He 等(2009)^[12]提出如下的迭代公式:

$$\beta^{(l+1)} = \beta^{(l)} + \{nA(\beta^{(l)}) + n \sum(\beta^{(l)})\}^{-1} Q(\beta) \quad (4)$$

对于给定的初值 $\beta^{(0)} = (\beta_1^{(0)}, \dots, \beta_d^{(0)})'$, 为了使惩罚函数在 0 点处可导, 我们有如下的线性近似:

$$p_{\lambda_i}'(|\beta_j|) \text{sgn}(\beta_j) \approx \{p_{\lambda_i}'(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\} \beta_j$$

其中, β_j 为第 l 次迭代 $\beta^{(l)}$ 的第 j 个分量。为了得到调整参数, 我们在以上每一次迭代中都使用广义交叉验证方法, 其定义为:

$$e(\lambda_1, \dots, \lambda_d) = \text{tr}[\{A(\beta) + \sum(\beta)\}^{-1} A(\beta)]$$

其中, tr 为求矩阵的迹, 即矩阵主对角线元素之和。则调整参数 $(\lambda_1, \dots, \lambda_d)$ 可以定义为如下统计量的最小值:

$$GCV(\lambda_1, \dots, \lambda_d) = \frac{\sum_{i=1}^n N_i \exp(-X_i^T \beta)}{n\{1 - e(\lambda_1, \dots, \lambda_d)/n\}^2}$$

将初值 $\beta^{(0)}$ 带入 GCV, 在得到 $\lambda_1, \lambda_2, \dots, \lambda_d$ 后, 即可计算 $A(\beta)$ 与 $\sum(\beta)$, 再带入(4)式, 迭代至收敛, 即可得到 β 的惩罚估计。

Tong and He 等(2009)^[12]指出, 在 d 维欧式空间上直接求解 λ 是非常困难的。为了能有效地得到 λ 值, 我们希望将该极值问题放到一维空间中求解。Li and Liang(2008)^[13]提出了富有启发的想法, 认为调整参数 λ_j 应当与无惩罚估计 $\tilde{\beta}_j$ 的标准差成正比, 此处 $\tilde{\beta}$ 为使 $W(\beta) = 0$ 的解, $\tilde{\beta}_j$ 为 $\tilde{\beta}$ 的第 j 个分量。以此为基础, Tong and He(2009)^[12]提出如下的定义:

$$\lambda = \lambda_0(\{SE(\tilde{\beta}_1)/|\tilde{\beta}_1|, \{SE(\tilde{\beta}_2)/|\tilde{\beta}_2|, \dots, \{SE(\tilde{\beta}_d)/|\tilde{\beta}_d|\}^T$$

其中, $SE(\hat{\beta}_j)$ 为求 $\hat{\beta}_j$ 的标准差。从上式可以看出, λ 中只含有一个未知参数 λ_0 , 将 λ 带入(5)式, 在一维空间中使 GCV 统计量达到最小, 得到 λ_0 , 进而得到调整参数 λ 。再计算 $A(\beta)$ 与 $\sum (\beta)$, 利用迭代公式(4), 使相邻两次迭代值之间的差异充分小, 即可得到 β 的估计值。

四、数值模拟

(一) 维数为6的数值模拟

利用模型 $E(N_i(t) | X_i) = \Lambda_0(t) \exp(X_i^T \beta)$ 产生 n 个数据点, 其中 X_i 服从标准正态分布, 维数为 6, $\beta = (1, 0, 1, 1, 0, 0)^T$, 删失时间 Y_i 服从参数为 10 的指数分布, 观测终止时间 τ 为无穷, 则 $t = Y_i$, 取 $\Lambda_0(t) = 0.1t$, N_i 取自均值为 $\Lambda_0(t) \exp(X_i^T \beta)$ 的泊松分布。

利用上节描述的惩罚估计方法, 将样本量分别取为 200、400、600, 当 $|\hat{\beta}_j| \leq 0.01$ 时, 我们令 $\hat{\beta}_j = 0$, 重复模拟 100 次, 得到的 β 估计见表 1。从表 1 可以看出, 随着样本量的增加, β 的估计越来越接近真值。

表 1 β 的估计值

n	β_1	β_2	β_3	β_4	β_5	β_6
200	0.9249	0	0.9825	0.8900	-0.0378	0.0492
400	0.9717	0.0513	0.9705	0.9714	0	0
600	1.0280	0	1.0576	1.0429	0	-0.0458

(二) 维数为10的数值模拟

利用模型 $E(N_i(t) | X_i) = \Lambda_0(t) \exp(X_i^T \beta)$ 产生 n 个数据点, 其中 X_i 服从 $[0, 1]$ 上的均匀分布, 维数为 10, 不失一般性, 我们假定已经对 X 进行了标准化处理, 令 $\beta = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)^T$, 删失时间 Y_i 服从参数为 10 的指数分布, 观测终止时间 τ 为无穷, 则 $t = Y_i$, 取 $\Lambda_0(t) = 0.5t^2$, N_i 取自均值为 $\Lambda_0(t) \exp(X_i^T \beta)$ 的泊松分布。

利用惩罚估计方法, 将样本量分别取 200、400、600, 当 $|\hat{\beta}_j| \leq 0.01$ 时, 我们令 $\hat{\beta}_j = 0$, 将得到的 β 估计见表 2, 上述过程重复模拟 100 次。由表 2 看出, 随着样本量的增加, β 的估计越来越接近真值。

表 2 β 的估计值

n	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
200	1.0837	0.9838	0.9064	0	-0.1117	0	0	0	0	0
400	1.0746	0.9676	1.0283	0	0	-0.1069	0	0	0	0
600	1.0676	1.0149	0.9823	0	0	0	0	0	0	0

同时我们利用估计的 $\hat{\beta}$ 与真实 β 之间的相关系数 $R^2(\beta)$ 来评价估计的贴近程度, R^2 越接近 1, 说明我们估计的效果越好。当样本量为 200、400、600 时, 重复 100 次模拟后 β 的平均估计值与真值之间的相关系数分别为 0.9897、0.9942、0.9971, 这说明随着样本量的增加, 估计的精度在提高, 惩罚估计方法得到的效果比较理想。

五、实例分析

本节基于澳大利亚 MAA 公司(TheMotor Accidents Authority)的一组综合险的索赔数据, 研究车

险索赔频数对影响因素的响应关系。该组数据共含有 1446 位投保人在 1993 年度的索赔信息, Jong and Heller(2008)^[14]利用 Copula 模型分析过该组数据。本文将利用模型(3)再次来分析该组数据, 我们将通过惩罚函数来挑选显著性因数变量, 从而达到降维目的。

以一份汽车保险合同在一个固定保险期内(一个保险期)的最终索赔频数为因变量, 影响因素为所有可能的变量, 共 17 个变量。几个比较重要的变量如下:

- (1)被保险人在该保险合同以前(不包括该保险合同期内的)的索赔金额;
- (2)被保险人性别(0 表示男性, 1 表示女性);
- (3)保单维持期(以一年为一个保单合同期, 表示被保险人在保险公司的合同连续维持了几年);
- (4)婚否(即被保险人在观测期内是否结婚, 0 表示未婚, 1 表示已婚);
- (5)父母健在(0 表示父母去世, 1 表示父母健在);
- (6)居住时间(以年为单位, 表示被保险人在同一处所居住的最长时间);
- (7)延误(即处理完索赔的耽误时间);
- (8)观测期数(即连续观测了多少时间, 以年为单位, 一年为一期)。

为了消除变量量纲的差异, 我们对所有协变量进行了标准化处理。利用惩罚估计方法对协变量进行降维, 得到系数 β 的估计 $\hat{\beta}$, 当 $|\hat{\beta}_j| \leq 0.01$ 时, 我们令 $\hat{\beta}_j = 0$, $\hat{\beta}$ 的值见表 4。

由表 4 可以看出, 对索赔频数影响较大的变量主要有: 前期的索赔金额、被保险人性别、被保险人婚否、被保险人父母是否健在以及最高受教育程度。通过以上估计, 我们可以得到如下结论:

(1)前期的索赔金额。在变量选择得到的系数估计中, 前期索赔金额的系数为 0.1658, 由此我们可知, 该变量对索赔频数有显著性影响。这主要是由于, 在正常情况下, 被保险人在过去的行为会自觉延续到现在, 这与行为经济学的基本假设相吻合。

(2)被保险人性别。在变量选择得到的系数估计中, 性别的系数为 0.5047, 这说明性别对交通事故的发生有较为显著的影响, 这主要是由于男女性格差异、行为模式等的不同造成男女在交通事故的发生次数及严重程度上有明显的区别。

(3)被保险人是否结婚。在表 4 中, 婚否对索赔频数的影响系数为 0.4783, 这说明是否结婚对因变量有较为显著的影响, 这主要是由于结婚使被保险者的家庭责任感上升, 从而自觉遵守交通规则, 减少交通事故的发生以及汽车保险的索赔次数。

(4)父母健在。从惩罚估计的结果来看, 该变量的系数估计为 0.2167, 这与我们直观上的感觉并不一致, 同被保险人是否已婚相同, 这主要是因为父母的健在使被保险人有更多的归属感及家庭责任感, 从而影响到交通事故的发生及汽车保险的索赔。

(6)最高受教育程度。从表 4 可以看出, 最高受教育程度对索赔次数有非常显著的影响。这是因为随着受教育程度的提高, 更高素质的被保险人会更加自觉地遵守交通法规, 从而对索赔次数的减少产生积极的影响。

(5)保单维持期。理论上, 保单维持期越长, 索赔次数越大, 保险理赔越高。然而, 被保险人性别、婚否、受教育程度等也对索赔次数有很大的影响, 从而使得保单维持期对索赔次数的影响不是那么显著。另外一个解释是, 由于汽车保险奖惩系统(Bonus-Malus System-BMS)的存在, 留在同一保险公司的长期客户都是“表现良好”的客户。

表 4 系数估计 $\hat{\beta}$ 的值

变量	数值	变量	数值
维持期	0	收入	0
婚否	0.4783	居住地区	0
旧索赔金	0.1658	性别	0.5047
索赔金额	0	受教程度	-0.4656
延误	0	职业	0
孩子数目	0	父母健在	0
年龄	0	家庭资产	0
居住时间	0	观测期数	0
职业年限	0		

六、总 结

本文中，我们提议了一个汽车保险索赔频数的均值计数模型，该模型允许每次索赔具体发生时间点缺失，同时也允许有高维协变量的存在。该方法无论从模型建立还是统计方法上讲都更具一般性和灵活性。我们利用凸惩罚变量选择方法对高维协变量进行降维，得到回归系数的稀疏估计，该方法提供了一种处理高维情形下车险索赔数据的另外一种选择。在本文中，我们主要研究了汽车保险的索赔次数，而没有考虑每次索赔的具体金额，这将是我们要继续研究的问题。

参考文献：

- [1] Lin D Y. Linear regression analysis of censored medical costs [J]. *Biostatistics*, 2000, 1(1): 35–47.
- [2] Lin D Y. Regression analysis of incomplete medical cost data [J]. *Statistics in Medicine*, 2003, 22(7): 1181–1200.
- [3] 赵晓兵, 王伟伟. 高维附加信息下的商业医疗保险费用评估模型和方法 [J]. *财经论丛*, 2013, (4): 58–65.
- [4] Faraway J. *Extending the Linear Model with R* [M]. Chapman & Hall/CRC, 2006.
- [5] Werner G., Modlin C. *Basic Ratemaking* [M]. Casualty Actuarial Society, 2010.
- [6] 孟生旺. 广义线性模型在汽车保险定价中的应用 [J]. *数理统计与管理*, 2007, (1): 24–29.
- [7] Wang MC, Qin J and Chiang CT. Analyzing recurrent event data with informative censoring [J]. *Journal of the American Statistical Association*, 2001, (96): 455–464.
- [8] Huang CY, Wang MC. Joint modeling and estimation of recurrent event processes and failure time data [J]. *Journal of the American Statistical Association*, 2004, (99): 1153–1165.
- [9] Zhao XB, Zhou X. Sufficient dimension reduction on the mean and rate functions of recurrent events [J]. *Statistics in Medicine*, 2014, 33(21), 3693–3709.
- [10] Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. *Journal of the American Statistical Association*, 2001, (96): 1348–1360.
- [11] Sun JG, Wei L. Regression analysis of panel count data with covariate-dependent observation and censoring times [J]. *Journal of the Royal Statistical Society: Series B*, 2000, (62): 293–302.
- [12] Tong XW, He X, Sun LQ, Sun JG. Variable selection for panel count data via non-concave penalized estimating function [J]. *Scandinavian Journal of Statistics*, 2009, (36): 620–635.
- [13] Li RZ, Liang, H. Variable selection in semiparametric regression modeling [J]. *The Annals of Statistics*, 2008, (36): 261–286.
- [14] Jong, P. and Heller, G. Z. *Generalized Linear Models for Insurance Data (International Series on Actuarial Science)* [M]. Cambridge, 2008.

Estimation of Car Insurance Claim Frequency under the Mean Count Model

ZHAO Xiao-bing, LIU Wei

(School of Mathematics & Statistics, Zhejiang University of Finance & Economics, Hangzhou 310018, China)

Abstract: Prediction of car insurance claim frequency is a focus of theoretical and empirical research of non-life actuarial studies. However, owing to the high-dimensional information involved, traditional models and estimation methods no longer apply. In this paper, some significant factors of car insurance claim frequency are identified through the variable selection method with convex penalty function based on the mean count model. A small simulation and a real data analysis are conducted to assess the feasibility of the proposed model and methods.

Key words: car insurance; mean count model; convex penalty; variable selection; estimate function

(责任编辑: 原 蕴)