Tweedie 分布在车险费率厘定中的应用

张连增 谢厚谊

(南开大学金融学院,天津 300350)

[摘 要] 在车险费率厘定中经常假设索赔频率与索赔强度分别服从泊松分布与伽玛分布,即假设总索赔服从复合泊松 - 伽玛分布。为了估计各风险类的纯保费(即总索赔均值),通常做法是对索赔频率与索赔强度分别建立广义线性模型(GLM),进而得到各风险类的索赔频率与索赔强度的均值,然后把两均值简单相乘即可;另一种做法利用复合泊松 - 伽玛分布是 Tweedie 分布的特例这一性质,直接对总索赔建立广义线性模型,进而也可以得到各风险类的总索赔均值。本文阐述了两种建模方法在处理车险费率厘定问题时的区别,通过对来自国外、国内的两组数据进行实证分析,比较了两种建模方法的优劣,并得到了一些初步结论。

[关键词] Tweedie 分布;复合泊松 - 伽玛分布;广义线性模型

[中图分类号] F840.63 [文献标识码] A [文章编号]1004 - 3306(2017)01 - 0080 - 11

DOI:10.13497/j.cnki.is.2017.01.008

一、引 言

近些年在非寿险定价的研究中,已有学者考察了索赔频率与索赔强度不相互独立的情形,并得出了很多有价值的结论(Shi 等,2015),但在很多情况下假设它们相互独立也是非常符合实际的,故本文依然假设它们是相互独立的。同时假设索赔频率与索赔强度分别服从泊松分布与伽玛分布,即假设总索赔服从复合泊松 - 伽玛分布。为了估计各风险类的纯保费(即总索赔均值),可以对索赔频率与索赔强度分别建立广义线性模型,进而得到各风险类的索赔频率与索赔强度的均值。由索赔频率与索赔强度的相互独立性,把两均值简单相乘即为各风险类的总索赔均值。

以上是保险公司通常的做法,也可使用另外一种做法:通过分析可知复合泊松 - 伽玛分布是 Tweedie 分布的特例,因而不妨直接对总索赔建立以 Tweedie 分布为响应变量的广义线性模型,进而也可以得到各风险类的总索赔均值。本文阐述了两种建模方法在处理车险费率厘定问题时的区别,通过对来自国外、国内的两组数据进行实证分析,比较了两种建模方法的优劣,并得到了一些初步结论。

本文第二节对指数分布族与广义线性模型进行了简单的回顾,重点介绍了 Tweedie 分布以及 Tweedie 分布与复合泊松 - 伽玛分布的关系;第三节阐述了两种建模方法各自的理论框架体系,并考察了两者的区别与联系;第四、五节用两种建模方法对国内外数据进行了分析与处理,比较了它们各自的结果,并得到了一些有用的结论。

二、广义线性模型与 Tweedie 分布

(一)指数分布族

广义线性模型的响应变量属于指数分布族,本节首先对指数分布族给出一个简单的回顾,更加详细的介

[基金项目] 本文得到国家自然科学基金(No.71271121)的资助。

[作者简介] 张连增,南开大学金融学院精算系教授,博士生导师,研究方向:精算与风险管理;谢厚谊,南开大学金融学院精算系硕士研究生,研究方向:精算与风险管理。

绍可参考有关专著(Jϕrgensen,1997)。

指数分布族是概率密度函数满足如下形式的一类分布:

$$f(y;\theta,\lambda)=c(y,\lambda)\exp(\lambda[\theta y-\kappa(\theta)]) \tag{1}$$

其中,θ 为典则参数,λ 为指示参数,$\kappa(\theta)$是一个光滑函数,$c(y,\lambda)$是使得式(1)满足规范化条件的非负函数。若随机变量 Y 的概率密度函数如式(1)所示,其矩母函数为:

$$m_Y(t)=E[e^{tY}]=\exp(\lambda[\kappa(\theta+t/\lambda)-\kappa(\theta)])$$

由矩母函数可得:

$$\mu=E[Y]=\kappa'(\theta)$$

$$Var[Y]=\kappa''(\theta)/\lambda$$

通过方差函数考察均值与方差的关系,方差函数定义为:

$$V(\mu)=\kappa''(\kappa'^{-1}(\mu))$$

则有:

$$Var[Y]=V(\mu)/\lambda$$

一些常见的分布均属于指数分布族。表 1 列出了几类常见的分布在指数分布族中的表示,以及它们的均值与方差函数(De Jong 和 Heller,2008)。

几类常见分布在指数分布族中的表示

表 1

分布	θ	λ	$\kappa(\theta)$	均值	方差函数
Normal (μ,σ^2)	μ	$1/\sigma^2$	$\theta^2/2$	μ	1
Poisson (μ)	$\ln(\mu)$	1	$\exp(\theta)$	μ	μ
Gamma (α,β)	$-\beta/\alpha$	α	$-\ln(-\theta)$	α/β	μ^2

(二)广义线性模型

广义线性模型在非寿险精算领域有着广泛的应用,目前已经成为对保险产品定价的常用模型。本节在保险产品定价的框架下简单介绍广义线性模型的相关概念与假设,更加详细的介绍可参考有关专著(Ohlsson 和 Johansson,2010)。

首先对保单进行风险分类,定义解释变量 $X_i,i=1,\cdots,k$,并由解释变量定义特性向量 $X=(X_1,X_2,\cdots,X_k)^T$,不同的特性向量对应不同的风险类。除了定义特性向量以外,还需给定权重,记为 W。例如,我们关注的是一年内的总索赔,则对于某保单三年内发生的平均每年总索赔,应取 $W=3$。

Y 为观测值随机变量,x、w 为 X、W 的一组给定数据,广义线性模型有如下假设:

1. 给定 $X=x$ 与 $W=w$,Y 的概率密度函数为

$$f_{Y|X,W}(y|x,w)=c(y,\lambda)\exp(\lambda[\theta y-\kappa(\theta)])$$

在广义线性模型中,Y 被称为响应变量,它属于指数分布族;

2. 指示参数 λ 的取值与 X 的取值无关,与 W 的取值有关:

$$\lambda=w/\varphi$$

其中,φ 被称为分散参数;

3. 给定 $X=x$,Y 的均值为:

$$E[Y]=h(x^T\beta)$$

其中,$\beta=(\beta_1,\beta_2,\cdots,\beta_k)^T$ 为系数向量,h 的反函数定义为连接函数 g,本文取 $g(\cdot)=\ln(\cdot)$。

在具体的广义线性模型建模过程中,由极大似然法可以得到 β 与 φ 的估计值。

(三)Tweedie 分布

Tweedie 分布属于指数分布族,本节简单介绍它的概念,更加详细的介绍可参考有关文献(Peters 等,2009)。

Tweedie 分布的方差函数具有如下特殊形式:

$V(\mu)=\mu^p$

其中,p 为常数。

当一个随机变量服从 Tweedie 分布时,$\kappa(\theta)$的函数表达式由 p 唯一确定。不妨选取均值 μ 代替典则参数 θ 作为新参数,则 Tweedie 分布可由一组参数(p,μ,λ)唯一确定,记为 $Tw(p,\mu,\lambda)$。例如,当 $p\in(1,2)$时,可以得到对应关系(Ohlsson 和 Johansson,2010):

$$\kappa(\theta)=-\frac{1}{p-2}[-(p-1)\theta]^{\frac{p-2}{p-1}}$$

$$\theta=-\frac{1}{p-1}\mu^{-(p-1)} \tag{2}$$

表 2 列出了当 p 取不同值时,Tweedie 分布对应的常见分布。由表 2 可知,当 $p\in(1,2)$时,Tweedie 分布即为复合泊松 - 伽玛分布。

不同 p 值下 Tweedie 分布对应的常见分布

表 2

p 值	分布	p 值	分布
0	正态分布	(1,2)	复合泊松 - 伽玛分布
1	泊松分布	2	伽玛分布

为了研究 Tweedie 分布与复合泊松 - 伽玛分布之间参数的相互关系,假设索赔频率 N 服从参数为 μ_N 的泊松分布,索赔强度 S 服从参数为(α,β)的伽玛分布,则总索赔 Y 服从参数为(μ_N,α,β)的复合泊松 - 伽玛分布,记为 $CPG(\mu_N,\alpha,\beta)$。通过对 Tweedie 分布与复合泊松 - 伽玛分布的矩母函数进行比较,可以得到对应关系(Jϕrgensen,1997):

$$CPG(\mu_N,\alpha,\beta)=Tw\left(\frac{\alpha+2}{\alpha+1},\frac{\mu_N\alpha}{\beta},\frac{\left(\frac{\mu_N\alpha}{\beta}\right)^{\frac{\alpha+2}{\alpha+1}-1}\beta}{\alpha+1}\right) \tag{3}$$

$$Tw(p,\mu,\lambda)=CPG\left(\frac{\lambda\mu^{2-p}}{2-p},-\frac{p-2}{p-1},\frac{\lambda\mu^{1-p}}{p-1}\right) \tag{4}$$

三、建模的理论框架

给定一组观测数据,包含若干个风险类,各风险类的索赔次数与索赔额数据已知。为了估计各风险类的纯保费(即总索赔 Y 的均值),有两种建模方法:一种是对索赔频率 N 与索赔强度 S 分别建立广义线性模型(其中索赔频率服从泊松分布,索赔强度服从伽玛分布),得到各风险类的索赔频率与索赔强度的均值 E[N]与 E[S],由于索赔频率与索赔强度相互独立,进而可以得到各风险类的总索赔 Y 的均值为索赔频率与索赔强度均值的简单相乘($E[Y]=E[N]\cdot E[S]$),我们把这种建模方法称为 SPG(Separate Poisson-Gamma)法;另一种建模方法直接对总索赔 Y 建立以 Tweedie 分布为响应变量的广义线性模型,进而也可以得到各风险类的总索赔 Y 的均值,我们把这种建模方法称为 Tweedie-GLM 法。

(一)SPG 法建模

实际应用 SPG 法建模时,可能会遇到索赔频率模型与索赔强度模型选取的解释变量不一致的情况,因

而需要对特性向量重新定义，使其包含所有的两类解释变量以满足风险分类的要求（Quijano 和 Garrido，2015）。为了讨论方便，这里不妨假设索赔频率与索赔强度模型选取的解释变量是一致的。

由极大似然法，可以得到索赔频率模型中系数向量 β^N 的估计值以及索赔强度模型中系数向量 β^S 与分散参数 φ_S 的估计值，进而第 i 风险类（$X = x_i$）的纯保费为：

$$E[Y^{(i)}] = E[N^{(i)}] \cdot E[S^{(i)}] = \exp(x_i^T\beta^N) \cdot \exp(x_i^T\beta^S) = \exp[x_i^T(\beta^N + \beta^S)] \tag{5}$$

其中，$N^{(i)}$、$S^{(i)}$ 与 $Y^{(i)}$ 分别为第 i 风险类的索赔频率、索赔强度与总索赔随机变量。

总索赔 $Y^{(i)}$ 服从复合泊松 - 伽玛分布，由表 1 可知，其参数为 $\mu_{N(i)} = \exp(x_i^T\beta^N)$，$\alpha = 1/\varphi_S$，$\beta^{(i)} = \alpha/\mu_{S(i)} = \exp(-x_i^T\beta^S)/\varphi_S$。进一步由式(3)可知，$Y^{(i)}$ 亦服从 Tweedie 分布，其参数为：

$$p = \frac{\alpha+2}{\alpha+1}, \mu^{(i)} = \frac{\mu_{N(i)}\alpha}{\beta^{(i)}}, \lambda^{(i)} = \frac{\left(\frac{\mu_{N(i)}\alpha}{\beta^{(i)}}\right)^{\frac{\alpha+2}{\alpha+1}-1}\beta^{(i)}}{\alpha+1} \tag{6}$$

可以看出 SPG 法建模得到的各风险类的总索赔 $Y^{(i)}$ 对应的 λ 值一般是不相等的。

（二）Tweedie-GLM 法建模

Tweedie-GLM 法直接对总索赔 Y 建立以 Tweedie 分布为响应变量的广义线性模型，具体实现过程可分为两步：

1. 为 Tweedie 分布选取一个合适的 p 值。由式(2)可知，这是为了确定 κ(θ) 的函数表达式，因为广义线性模型要求 κ(θ) 的函数表达式事先给定。可以由极大似然法求出 p 的最优估计值 $\hat{p}$，求解过程一般会用到数值计算。R 软件中的 Tweedie 软件包包含这一求解方法（Dunn，2016）。

2. 由极大似然法得到总索赔模型中系数向量 β^Y 与分散参数 φ_Y 的估计值。

则第 i 风险类的纯保费为：

$$E[Y^{(i)}] = \exp(x_i^T\beta^Y) \tag{7}$$

总索赔 $Y^{(i)}$ 服从 Tweedie 分布，其参数为：

$$p = \hat{p}, \mu^{(i)} = \exp(x_i^T\beta^Y), \lambda = 1/\varphi_Y \tag{8}$$

可以看出 Tweedie-GLM 法建模得到的各风险类的总索赔 $Y^{(i)}$ 对应的 λ 值是相等的，均等于 $1/\varphi_Y$。

（三）两种建模方法的比较

由以上分析可知，各风险类的总索赔 $Y^{(i)}$ 对应的 λ 值是否相等，两种建模方法给出了不同的结论，因此两种方法本质上是有区别的。

可以从另一侧面考察这种区别：由 Tweedie-GLM 法建立的总索赔模型导出其隐含的“索赔频率 - 索赔强度”复合模型，然后与 SPG 法建立的模型进行比较。由于推导过程繁琐，这里不再赘述（Quijano 和 Garrido，2015）。但即使不去关注这一隐含复合模型的具体形式，仅由式(4)就可以很方便的得出该隐含模型中各风险类的索赔频率与索赔强度的均值。第 i 风险类有：

$$\mu_{N(i)}^* = \frac{\lambda(\mu^{(i)})^{2-p}}{2-p}$$

$$\mu_{S(i)}^* = \frac{\left(-\frac{p-2}{p-1}\right)}{\left(\frac{\lambda(\mu^{(i)})^{1-p}}{p-1}\right)} = \frac{(2-p)(\mu^{(i)})^{p-1}}{\lambda} \tag{9}$$

其中，p、$\mu^{(i)}$ 与 λ 的值由式(8)给定，上标星号表示所求均值为隐含模型中索赔频率与索赔强度的均值，与 SPG 法建模得到的均值加以区分。

式(9)中各风险类的 λ 与 p 保持恒定，$\mu^{(i)}$ 随风险类的不同而改变。可以看出对不同的风险类，索赔频率的均值 $\mu_{N(i)}^*$ 与索赔强度的均值 $\mu_{S(i)}^*$ 同时增大或减小，这是因为 $\mu^{(i)}$ 的幂指数（2 - p）与（p - 1）均为大于 0 的数。由于 SPG 法建模并没有对索赔频率与索赔强度的均值是否要同时增大或减小做出任何要求，从这一

侧面也可知两种建模方法是有本质不同的。

因此,Tweedie-GLM 法建模对索赔频率与索赔强度有更为严格的假设。在实际应用中,对于给定的数据样本,它是否满足这种更为严格的假设并不影响我们选取 Tweedie-GLM 法对其进行建模。只要 Tweedie-GLM 法建立的模型对各风险类的纯保费的估计能够与观测数据相吻合,这种建模方法就是可取的。我们关心的是总索赔模型能否很好的估计出各风险类的纯保费,并不关心隐含的索赔频率或索赔强度模型对数据拟合的好坏。另外,相比于 SPG 法建模,Tweedie-GLM 法建模用到的待估参数一般更少,这一点在后面的实证分析中会得到印证。

四、国外数据实证分析

选取一个国外的具体实例进行分析,数据来源于瑞典 Wasa 保险公司的摩托车全损保险理赔数据①。这里的"全损"特指车辆自燃或被盗,并非每个国家都有该险种。表 3 列出了保单的各风险因子及其等级,并对各等级进行了定义,可以看出共有 $2\times2\times7=28$ 个风险类。

需要注意,本文使用的是各风险类的汇总索赔数据,并非每份保单的个体索赔数据。例如,车型(Class)等级为 2、车龄(Age)等级为 1、区域(Zone)等级为 6 时(简记此风险类为 C2A1Z6 类),风险暴露数为 82.8 车年,总的索赔次数为 3 次,3 次索赔的总额为 17490 瑞典克朗。由于某些风险类的风险暴露数太少(C1A1Z5、C1A1Z7、C2A1Z5 与 C2A1Z7 类),为确保统计性要求,本文对这些风险类予以剔除,剔除后还有 $28-4=24$ 个风险类。

摩托车全损保险风险分类

表 3

风险因子	等级	定义
车型	1	车重超过 60 千克且档位超过两个
	2	其他
车龄	1	≤1 年
	2	>1 年
区域	1	三大城市的中心及次中心区域
	2	三大城市的郊区及中等城市
	3	除 5、7 之外的次中等城市
	4	除 5~7 之外的小城市及乡村
	5	北部城市
	6	北部乡村
	7	哥特兰岛(瑞典最大的岛屿)

(一)SPG 法建模

SPG 法对索赔频率与索赔强度分别建立广义线性模型。对索赔频率建模时,选取各风险类的风险暴露数为 offset 项(De Jong 和 Heller,2008);对索赔强度建模时,选取各风险类的总的索赔次数为权重。

1. 索赔频率建模

对索赔频率建模,首先需要判断相比于泊松分布模型,负二项分布模型是否更为合适。两种不同模型的初步建模结果如图 1 与图 2 所示,它们被称为索赔频率均值对比图。图中横坐标代表 24 个不同的风险类;右侧纵坐标表示各风险类的风险暴露数;左侧纵坐标表示各风险类的索赔频率均值,x 线表示索赔频率均值

① 可访问 http://staff.math.su.se/esbj/GLMbook/moppe.sas 进行下载。

的观测值，o 线表示索赔频率均值的模型估计值。由于已按照索赔频率均值的模型估计值从小到大的顺序对各风险类进行了重新排序，o 线是保持向上倾斜的。通过比较可以看出，泊松分布模型对索赔频率均值的估计更接近于观测值，因而我们选择泊松分布模型对索赔频率建模。

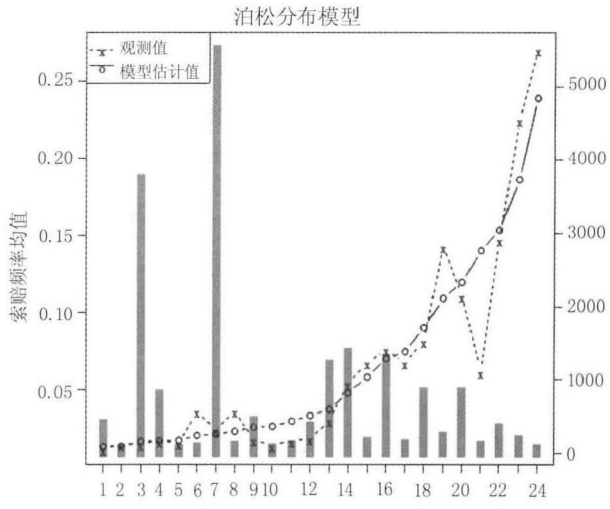

图 1　泊松分布模型的均值对比

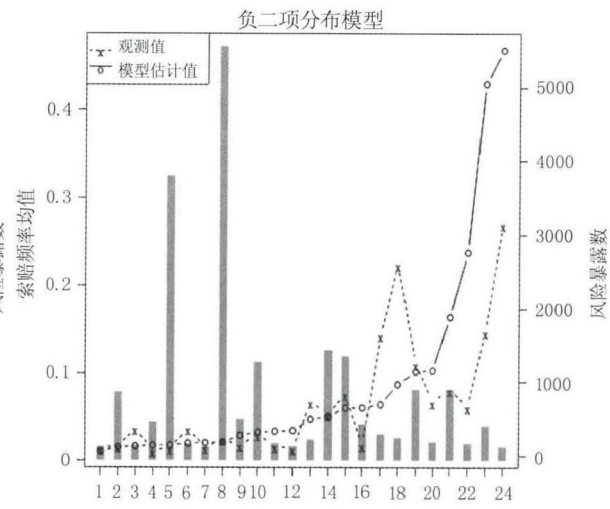

图 2　负二项分布模型的均值对比

通过对泊松分布模型的进一步分析，加入交互项 C1A1Z2TRUE 与 C1A2Z4TRUE 作为解释变量（C1A1Z2 类的 C1A1Z2TRUE 取值为 1，其余各风险类的 C1A1Z2TRUE 取值为 0；C1A2Z4TRUE 的含义类似），建模结果如表 4 所示。可以看出各解释变量均显著，并且由于我们研究的是摩托车自燃或被盗险，好车（车型等级为 1）与新车（车龄等级为 1）更容易被偷窃，Class2 与 Age2 的系数均应为负值，建模结果给出了相同的结论。卡方统计量的值为 11.17（Jϕrgensen，1992），自由度为 13，$Pr(\chi^2_{13} \geq 11.17) = 0.597$，因此不能拒绝泊松分布模型对索赔频率建模。

泊松分布模型对索赔频率建模结果

表 4

	估计值	标准误	z 值	Pr(> |z|)
(Intercept)	-1.38146	0.11794	-11.713	0.0000 ***
Class2	-0.19301	0.08448	-2.285	0.0223 *
Age2	-0.5597	0.09883	-5.663	0.0000 ***
Zone2	-0.46971	0.09975	-4.709	0.0000 ***
Zone3	-1.13014	0.11373	-9.937	0.0000 ***
Zone4	-2.18267	0.13839	-15.772	0.0000 ***
Zone5	-1.58666	0.4154	-3.82	0.0001 ***
Zone6	-2.15893	0.22195	-9.727	0.0000 ***
Zone7	-2.12755	0.71121	-2.991	0.0028 **
C1A1Z2TRUE	-0.92941	0.39779	-2.336	0.0195 *
C1A2Z4TRUE	0.42302	0.16483	2.566	0.0103 *

显著性标志：0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

（泊松分布分散参数值为：1）

零偏差：23 个自由度，值为 515.48

残余偏差：13 个自由度，值为 11.17

2. 索赔强度建模

通过对伽玛分布模型的进一步分析，只保留“区域”风险因子中的等级 1 作为解释变量（用 Z1TRUE 表示），建模结果如表 5 所示。可以看出各解释变量均显著，并且由于好车（车型等级为 1）与新车（车龄等级为 1）自燃或被盗后的损失更大，Class2 与 Age2 的系数均应为负值，建模结果给出了相同的结论。卡方统计量的值为 8.6675/0.4690377 = 18.48，自由度为 20，$Pr(\chi^2_{20} \geq 18.48) = 0.556$，因此不能拒绝伽玛分布模型对索赔强度建模。

伽玛分布模型对索赔强度建模结果

表 5

	估计值	标准误	t 值	Pr(> |t|)
(Intercept)	9.48777	0.07027	135.028	0.0000 ***
Class2	-0.59658	0.05049	-11.816	0.0000 ***
Age2	-0.59178	0.06573	-9.003	0.0000 ***
Z1TRUE	0.1466	0.05723	2.562	0.0186 *

显著性标志：0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

（伽玛分布分散参数值为 0.4690377）

零偏差：23 个自由度，值为 109.7342

残余偏差：20 个自由度，值为 8.6675

3. 纯保费估计

通过对索赔频率与索赔强度分别建模，得到了各风险类的索赔频率与索赔强度均值的模型估计值，把这两个估计值简单相乘，即为各风险类的总索赔均值的模型估计值，即纯保费的估计值。与图 1、图 2 类似，图 3 被称为总索赔均值对比图，图中横坐标代表 24 个不同的风险类；右侧纵坐标表示各风险类的风险暴露数；左侧纵坐标表示各风险类的总索赔均值，x 线表示总索赔均值的观测值，o 线表示总索赔均值的模型估计值。可以看出大多数情况下 SPG 法建立的模型对各风险类的纯保费的估计能够与观测值相吻合。图 4 为 Q - Q 图，可以看出偏离较大的点只有一个，对应图 3 中的第 24 风险类，由于此风险类的风险暴露数较小，同样可以得到模型能否很好的估计出各风险类的纯保费这一结论。

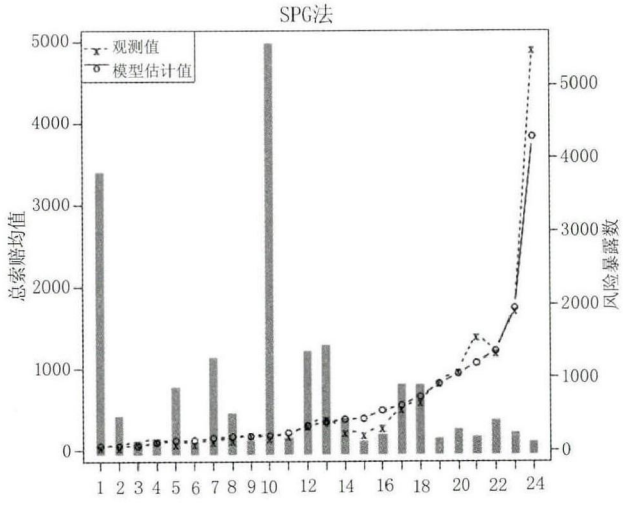

图 3　SPG 法总索赔均值对比

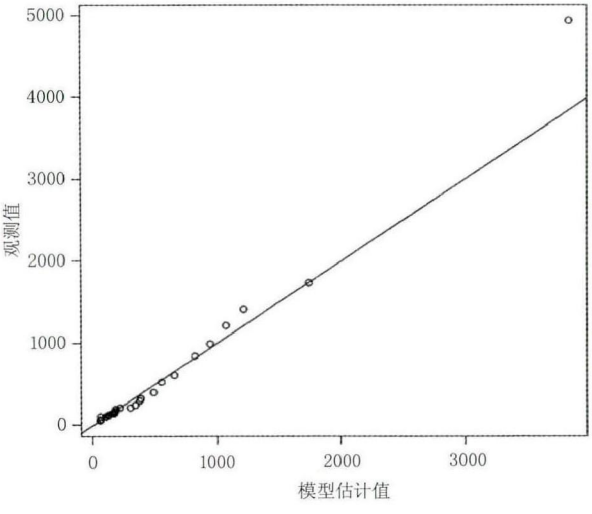

图 4　SPG 法 Q - Q 图

(二)Tweedie-GLM 法建模

不妨选择 SPG 法建模用到的所有解释变量作为 Tweedie-GLM 法建立的总索赔模型的解释变量。建模结果如表 6 所示。可以看出各解释变量均显著,卡方统计量的值为 502.54/38.11064 = 13.19,自由度为 13,$Pr(\chi^2_{13} \geq 13.19) = 0.434$,因此不能拒绝 Tweedie 分布模型对总索赔建模。

Tweedie 分布模型对总索赔建模结果

表 6

	估计值	标准误	t 值	Pr(> |t|)
(Intercept)	8.11911	0.16185	50.164	0.0000***
Class2	-0.75706	0.09654	-7.842	0.0000***
Age2	-1.02278	0.11884	-8.606	0.0000***
Zone2	-0.59016	0.14907	-3.959	0.0016**
Zone3	-1.25551	0.14687	-8.549	0.0000***
Zone4	-2.32318	0.13861	-16.761	0.0000***
Zone5	-1.58582	0.34111	-4.649	0.0005***
Zone6	-2.34941	0.17957	-13.084	0.0000***
Zone7	-1.92015	0.44479	-4.317	0.0008***
C1A1Z2TRUE	-0.78936	0.44131	-1.789	0.0970
C1A2Z4TRUE	0.3685	0.13451	2.74	0.0169*

显著性标志:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Tweedie 分布分散参数值为 38.11064)

零偏差:23 个自由度,值为 28305.32

残余偏差:13 个自由度,值为 502.54

通过对总索赔建模,可以得到各风险类的总索赔均值的模型估计值。对比总索赔均值的模型估计值与观测值(如图 5 所示),可以看出大多数情况下 Tweedie-GLM 法建立的模型对各风险类的纯保费的估计能够与观测值相吻合。图 6 为 Q-Q 图,可以看出偏离较大的点只有一个,对应图 5 中的第 24 风险类,由于此风险类的风险暴露数较小,同样可以得到模型能否很好的估计出各风险类的纯保费这一结论。

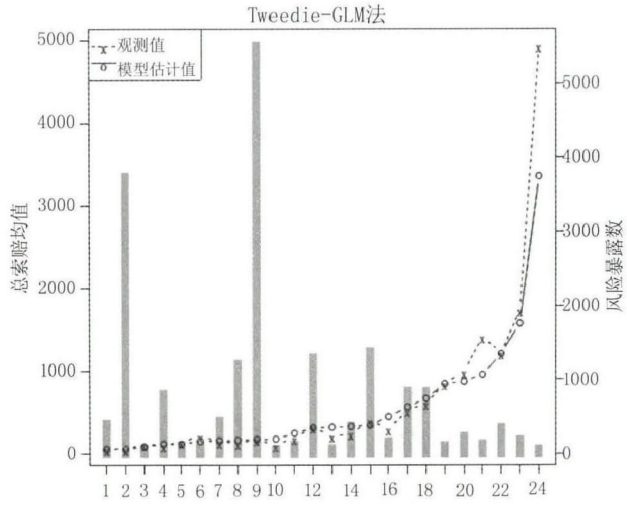

图 5 Tweedie-GLM 法总索赔均值对比

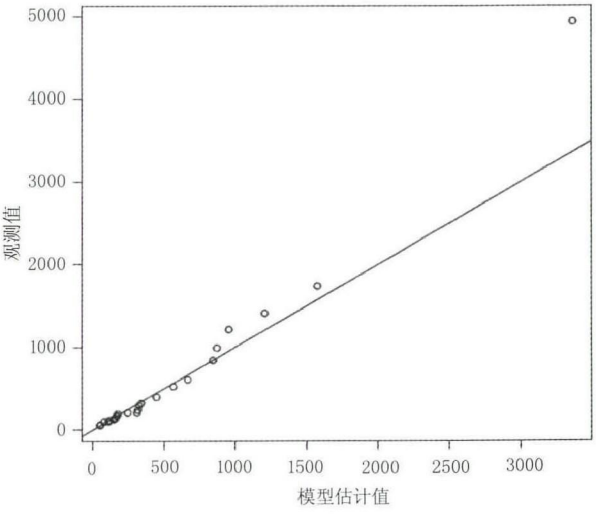

图 6 Tweedie-GLM 法 Q-Q 图

(三)模型比较

比较图 3 与图 5,可以看出两种建模方法均能够很好的估计出各风险类的纯保费。SPG 法建模用到了

11 +5 =16 个待估参数(索赔频率模型有 11 个待估参数,索赔强度模型有 4 +1 =5 个待估参数,其中 1 个是分散参数 φ_S);Tweedie-GLM 法建模用到了 11 +1 =12 个待估参数(其中 1 个是分散参数 φ_Y)。因此,从待估参数个数的角度看,Tweedie-GLM 法建模更优。

由 Tweedie-GLM 法建立的总索赔模型可以导出其隐含的"索赔频率 - 索赔强度"复合模型,该隐含模型对数据拟合的并不好(如图 7 与图 8 所示),但不能因此而拒绝使用 Tweedie-GLM 法建模。因为我们关心的是总索赔模型能否很好的估计出各风险类的纯保费,并不关心隐含的索赔频率或索赔强度模型对数据拟合的好坏。只要 Tweedie-GLM 法建立的模型对各风险类的纯保费的估计能够与观测数据相吻合,就应该优先选择 Tweedie-GLM 法建模,因为相比于 SPG 法建模,Tweedie-GLM 法建模一般用到更少的待估参数。

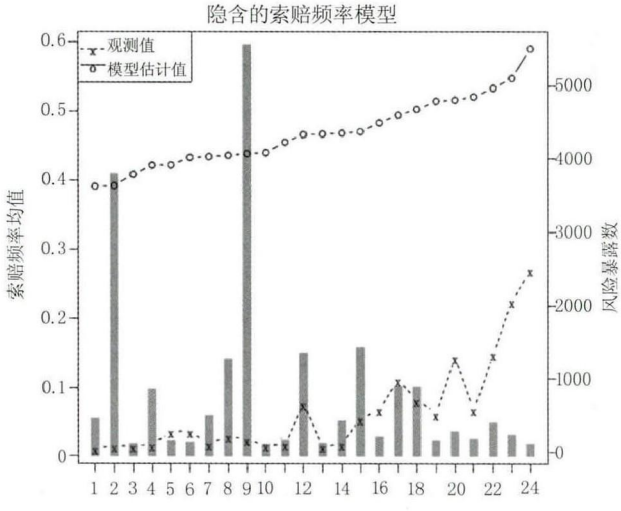

图 7　隐含的索赔频率模型均值对比

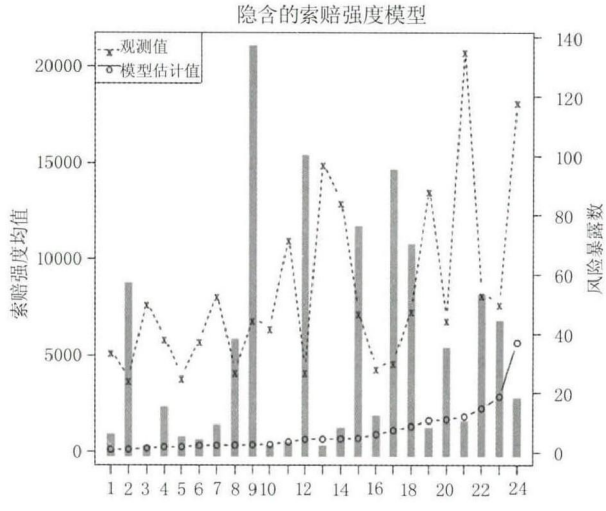

图 8　隐含的索赔强度模型均值对比

五、国内数据实证分析

选取国内某保险公司在某省份转保的 6 座以下某车系家庭自用车交通事故责任强制保险理赔数据进行分析。表 7 列出了保单的各风险因子及其等级,并对各等级进行了定义。可以看出共有三个风险因子,分别为性别(Sex)、年龄(Age)与无赔款优待(Class)。为了简化问题,只研究 25 ~ 50 岁之间无赔款优待处于前三个等级的保单,因此共有 2 ×5 ×3 =30 个风险类。这 30 个风险类的汇总数据来自 26502 条原始数据记录。

家庭自用车交通事故责任强制保险风险分类

表 7

风险因子	等级	定义
性别	1	男性
	2	女性
年龄	1	25 ~ 30 岁
	2	30 ~ 35 岁
	3	35 ~ 40 岁
	4	40 ~ 45 岁
	5	45 ~ 50 岁
无赔款优待	1	连续三年及以上未发生有责任事故
	2	连续两年未发生有责任事故
	3	上一年未发生有责任事故

与第四节分析摩托车全损保险理赔数据的方法类似，首先用 SPG 法建模，其中对索赔频率建模时加入了交互项 S1A2C2TRUE、S2A1C1TRUE 与 S2A3C1TRUE 作为解释变量，对索赔强度建模时加入了交互项 S1A4C2TRUE、S1A2C1TRUE 与 S2A1C2TRUE 作为解释变量。然后用 Tweedie-GLM 法建模，并且选择 SPG 法建模用到的所有解释变量作为总索赔模型的解释变量。

两种建模方法得到的总索赔均值的模型估计值如图 9 所示，图中横坐标代表 30 个不同的风险类；右侧纵坐标表示各风险类的风险暴露数；左侧纵坐标表示各风险类的总索赔均值的模型估计值，o 线表示 SPG 法建模得到的总索赔均值的模型估计值，x 线表示 Tweedie-GLM 法建模得到的总索赔均值的模型估计值。由于已按照 SPG 法建模得到的总索赔均值的模型估计值从小到大的顺序对各风险类进行了重新排序，o 线是保持向上倾斜的。可以看出大多数情况下两种建模方法建立的模型对各风险类的纯保费的估计是相近的。

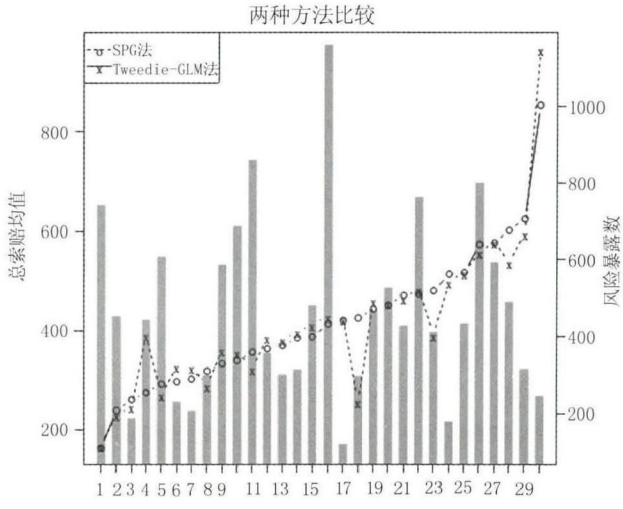

图 9　总索赔均值对比

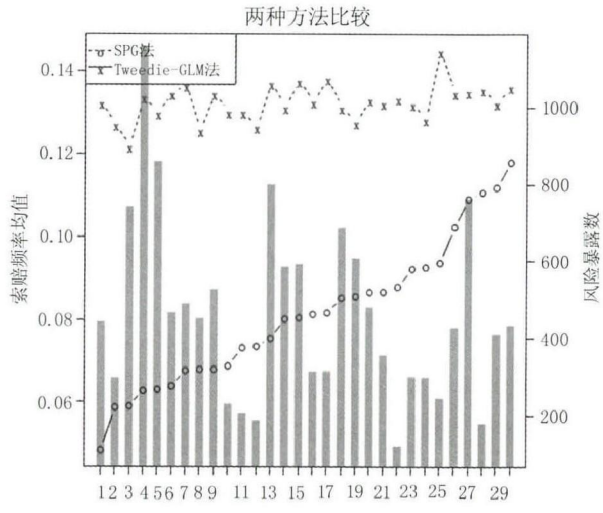

图 10　索赔频率均值对比

由于 SPG 法建模用到了 11 + 12 = 23 个待估参数（索赔频率模型有 11 个待估参数，索赔强度模型有 11 + 1 = 12 个待估参数，其中 1 个是分散参数 φ_S），Tweedie-GLM 法建模用到了 14 + 1 = 15 个待估参数（其中 1 个是分散参数 φ_Y），从待估参数个数的角度看，Tweedie-GLM 法建模更优。

由 Tweedie-GLM 法建立的总索赔模型可以导出其隐含的“索赔频率 - 索赔强度”复合模型，图 10 中 x 线表示隐含的索赔频率模型得到的索赔频率均值的模型估计值，o 线表示 SPG 法建模得到的索赔频率均值的模型估计值。由于已按照 SPG 法建模得到的索赔频率均值的模型估计值从小到大的顺序对各风险类进行了重新排序，o 线是保持向上倾斜的。可以看出两种建模方法建立的模型对各风险类的索赔频率均值的估计相差甚远，但不能因此而拒绝使用 Tweedie-GLM 法建模，因为我们关心的是总索赔模型，并不关心隐含的索赔频率或索赔强度模型。

六、结　论

本文通过两个具体的保险产品定价实例，比较了车险费率厘定问题的 SPG 法建模与 Tweedie-GLM 法建模。当两种建模方法均能够很好的估计出各风险类的纯保费时，应该优先选择 Tweedie-GLM 法建模，因为相比于 SPG 法建模，Tweedie-GLM 法建模一般用到更少的待估参数。

本文是基于各风险类的汇总数据进行建模的①，当使用的是个体数据而不是汇总数据时，Tweedie-GLM 法建立的模型是否还能够很好的估计出各风险类的纯保费，需要进一步的研究分析。

① 本文所有的运算均由 R 软件完成。读者如对相应的数据与代码感兴趣，可联系作者。

[参考文献]

[1] De Jong P, Heller GZ. Generalized linear models for insurance data [M]. New York: Cambridge University Press, 2008.

[2] Dunn PK. Tweedie: Tweedie exponential family models. R package. Version 2.2.5.2016.

[3] Jϕrgensen B, Paes de Souza MC. Fitting Tweedie's compound Poisson model to insurance claims data [J]. Scandinavian Actuarial Journal, 1994, (1): 69 - 93.

[4] Jϕrgensen B. The theory of dispersion models [M]. London: Chapman & Hall, 1997.

[5] Jϕrgensen B. The theory of exponential dispersion models and analysis of deviance [M]. Brazil: Instituto de Matemática Pura e Aplicada, 1992.

[6] Ohlsson E, Johansson B. Non-life insurance pricing with generalized linear models [M]. Berlin: Springer, 2010.

[7] Peters GW, Shevchenko PV, Wüthrich MV. Model uncertainty in claims reserving within Tweedie's compound Poisson models [J]. Astin Bulletin, 2009, 39(1): 1 - 33.

[8] Quijano Xacur OA, Garrido J. Generalised linear models for aggregate claims: to Tweedie or not? [J]. European Actuarial Journal, 2015, 5(1): 181 - 202.

[9] Shi P, Feng X, Ivantsova A. Dependent frequency-severity modeling of insurance claims [J]. Insurance: Mathematics and Economics, 2015, 64: 417 - 428.

[10] Smyth GK, Jϕrgensen B. Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling [J]. Astin Bulletin, 2002, 32(1): 143 - 157.

[11] Smyth GK, Verbyla AP. Adjusted likelihood methods for modelling dispersion in generalized linear models [J]. Environmetrics, 1999, 10(6): 695 - 709.

The Application of Tweedie Distribution in Auto Insurance Ratemaking

ZHANG Lianzeng, XIE Houyi

(Department of Actuarial Science, School of Finance, Nankai University, Tianjin 300350)

Abstract: Auto insurance rate-making often assumes the claim frequency and claim severity follows Poisson distribution and Gamma distribution respectively, namely, the total claim follows compound Poisson-Gamma distribution. Under this distributional assumption, generalized linear models (GLMs) are used to estimate the mean claim frequency and severity, then these estimators are simply multiplied to estimate the net premiums or the mean aggregate loss for various risks. Because the compound Poisson-Gamma distribution is a Tweedie distribution in nature, therefore another method is to establish a GLM for the total claims, and then arrive at the mean of total claims for various risks. The paper elaborated on the differences of these two modeling methods for auto insurance ratemaking. Through an exponential analysis of two data sets, both international and domestic, it compared their advantages and disadvantages and offered some initial conclusions.

Key words: Tweedie distribution; CPG distribution; generalized linear models

[编辑:施　敏]