文章编号: 1002 - 1566(2007)01 - 0024 - 06

广义线性模型在汽车保险定价的应用

孟生旺

(中国人民大学统计科学研究中心, 100872)

摘要:对非寿险产品分类费率的厘定通常采用单项分析法、最小偏差法和多元线性回归等方法。虽然这些方法在非寿险产品定价中仍然占有一度之地,但由于保险数据的特殊性,它们的缺陷越来越受到人们的重视。本文简要分析了这些传统定价方法存在的缺陷,介绍了非寿险精算中典型的广义线性模型,并通过汽车第三者责任保险的损失数据说明了广义线性模型在非寿险产品定价中的具体应用,以及应用广义线性模型时应该注意的几个问题。

关键词:广义线性模型;汽车保险;定价

中图分类号: O212 **文献标识码**: A

An Application of Generalized Linear Model to Automotor Insurance Pricing

MENG Sheng-wang

(Statistics Research Center, Renmin University of China, Beijing 100872)

Abstract: One-way analysis, minimum bias procedures and multiple regressions have been widely used in non-life classification rating. But their drawbacks are becoming very clear when considering properties of insurance data. The paper analyzes the drawbacks of these traditional methods, discusses the application of generalized linear models in non-life insurance pricing, and points out some problems to be considered when applying generalized linear models.

Key Words: generalized linear models; automobile insurance; pricing

1 传统定价方法的局限性

1.1 单项分析法

单项分析法根据每个费率因子分别确定其对保险产品价格的影响。单项分析法由于不能考虑各个费率因子之间的相互关系,很容易导致保险价格的扭曲。譬如,在汽车保险中,对车龄的单项分析结果表明,汽车越旧,其保险成本越高。但导致这种现象的真正原因可主要是因为高风险的年轻人驾驶旧车的可能性较大,才导致了旧车的保险成本较高。因此,如果根据车龄和驾驶员年龄的单项分析结果厘定汽车保险费率,将会重复使用驾驶员年龄对汽车保险费率的影响,最终导致对年轻驾驶员收取了过高的保险费。

单项分析法的另一个缺陷是没有考虑各个费率因子之间的相互依存关系或交互作用。譬如,男女驾驶员的保险成本差异通常随驾驶员年龄的变化而变化,这就表明驾驶员的性别和年龄之间存在交互作用,但单向分析法难以区分这种差异,而会简单地认为男女驾驶员的保险成本差异与年龄无关。

收稿日期: 2005年 6月 17日

基金项目:教育部人文社会科学项目 (05JJD910152)

1.2 最小偏差法

最小偏差法是在 19世纪 60年代发展起来的一种分类费率厘定方法,该方法通过一个方程组建立损失数据和各个费率因子之间的关系,并通过迭代法求解未知参数的最优解。与单项分析法相比,最小偏差法有了很大进步,但最优解一旦确定以后,最小偏差法并不能提供一种统计方法对特定费率因子的显著性进行检验,也不能确定参数估计的置信区间。因此,最小偏差法的主要缺陷是缺乏一个完整的统计分析框架对建模结果进行评价。

1.3 多元线性回归模型

多元线性回归模型在非寿险分类费率的厘定中有很广泛的应用,但其严格的假调条件在非寿险中通常难以得到满足;

首先,要求因变量服从正态分布在很多情况下是不现实的,譬如索赔频率和续保率等通常不会服从正态分布。

其次,非寿险的因变量(如索赔频率和次均赔款等)通常是非负的,而正态分布的假设显然不能满足这一要求。

第三,如果因变量是严格非负的,那么从直观上看,当因变量的均值趋于零时,其方差也应该趋于零,即因变量的方差应该是其均值的函数。但在多元线性回归模型中,假设因变量的方差是固定的常数,与均值没有任何关系。

第四,在多元线性回归模型中,假设费率因子通过加法关系对因变量产生影响,但在很多情况下,费率因子之间可能是一种乘法关系,而非加法关系。

2 非寿险精算中典型的广义线性模型

与传统的线性回归模型一样,广义线性模型也基于一系列假设之上,但这些假设要相对宽松很多。广义线性模型的假设由随机成分、系统成分和联结函数三部分组成的:

(1)随机成分,即因变量 Y或误差项的概率分布。因变量 Y的每个观察值 y_i相互独立服从指数型分布族中的一种分布。指数型分布族包括许多常见分布,如正态分布、泊松分布、逆高斯分布、二项分布、伽玛分布等。

(2)系统成分,即自变量的线性组合,表示为 $\eta=\beta_1 x_1+\cdots+\beta_p x_p$ 或 $X\beta$。系统成分与多元线性回归模型没有任何区别。

(3)联结函数。联结函数 g单调且可导,它建立了随机成分与系统成分之间的关系,即 $E[Y]=\mu=g^{-1}(\eta)$。可见,广义线性模型对因变量的预测值并不直接等于自变量的线性组合,而是该线性组合的一个函数变换。

不难看出,传统的线性回归模型是广义模型是广义线性模型的特例。此外,可以证明(参见文献 2),许多最小偏差模型也是广义线性模型的特例。

在广义线性模型中,因变量的方差是其均值的函数,这一特点非常适合于保险数据。譬如在拟合次均赔款时,如果使用传统的线性回归模型,假设误差项服从正态分布,那么当拟合值为 100元时,标准误差为 10元,而拟合值为 10000元时,标准误差也为 10元。但从直观上看,拟合值越大,其标准误差也应越大。在广义线性模型中,如果假设因变量的变异系数为常数(如 10%)。这就就意味着假设误差项服从伽玛分布,在这种情况下,当拟合值为 100元时,标准误差为 10元,而当拟合值为 10000元时,标准误差为 1000元。

在广义线性模型的应用中,根据保险数据的先验知识选择误差项的分布类型,可以有效改

进模型的拟合效果。譬如,如果因变量的方差为常数,可以选择正态分布;如果因变量的方差等于其均值,可以选择泊松分布;如果因变量的方差等于其均值的平方,可以选择伽玛分布;如果因变量的方差等于其均值的三次方,可以选择逆高斯分布。

在非寿险产品的定价中,通常需要估计索赔次数、索赔频率、次均赔款和续保率等,根据这些数据的特点,各种典型的广义性模型分别是:

(1)在估计索赔次数或索赔频率时,典型的广义线性模型是泊松乘法模型,即使用对数联结函数和泊松分布的误差项。用泊松分布描述索赔次数时,不会受到时间度量单位的影响,无论是使用每月的索赔次数还是每年的索赔次数,结果是一致的。在估计索赔次数时,通常用风险单位数的对数作偏移项,而在估计索赔频率时,通常用风险单位数加权。

(2)在估计次均赔款时,典型广义线性模型是伽玛乘法模型,即使用对数联结函数和伽玛分布的误差项。伽玛分布不受货币量单位的影响,无论使用人民币还是美元作为度量单位,伽玛乘法广义线性模型的结果不会改变。

(3)在估计续保率和新业务转换率时,典型的广义线性模型是 Logistic模型,即采用分布对数 (Logit)联结函数和二项分布的误差项。分对数联结函数建立了从 (0, 1)到 (- ∞, + ∞)的映射关系,而且不论采用失败率还是成功率,其结果不变。如果该模型主要用于定性分析而非定量分析,则当因变量的取值很小时,可以用泊松乘法广义线性模型近似。

3 在汽车保险定价中的应用

下面用汽车第三者责任保险的一组损失数据讨论广义线性模型在保险费率厘定中的具体应用 (数据来源:WWW. StatSci org)。该数据包含 7个变量:年行驶里程数 (分为 5个等级,用 K表示),无赔款折扣等级 (分为 7个等级,用 B表示),行驶地区 (分为 7类,用 Z表示),车型 (分为 9类,用 M表示),保单年数,索赔次数和赔付额;所有的被保险人被划分成了 2182个类别 (应该为 5 ×7 ×7 ×9 = 2205个类别,其中 23个类别没有被保险人),总的保单年数为 2383170. 08。经验数据的平均索赔频率为 0. 04749,次均赔款为 4955. 25,平均纯保费为 235. 31。

在进行广义线性模型分析之前,我们首先用传统的多元线性回归模型对索赔频率和次均赔款数据进行了拟合,自变量是年行驶里程数、车型、行驶地区和无赔款折扣等级。结果表明,对索赔频率和次均赔款的拟合结果出现了负值,这显然是不合常理的。

如果采用典型的广义线性模型,假设索赔频率服从泊松分布,选择对数联结函数,用保单年数加权;假设次均赔款服从伽玛分布,选择对数联结函数,并用索赔次数加权,则调用 SAS/GENMOD对索赔频率,次均赔款和纯保费的拟合结果如表 1所示。

如果在拟合索赔频率时,使用正态分布假设和对数联结函数,在拟合次均赔款时,使用逆高斯假设和对数联结函数,则模型的整体拟合效果可以得到优化。有关比较结果如表 2和表 3所示。

不过,尽管将泊松假设改为正态假设,将伽玛假设改为逆高斯假设,模型的整体拟合优度得到了改善,但对纯保费的预测影响并不算大,两者预测值的平均绝对值差异只有 3. 12%。这表明典型模型对这组数据也是适用的。

表 1 索赔频率、次均赔款和纯保费的预测值

变量	索赔频率 (1)		次均赔款 (2)		纯保费 (3) = (1) * (2)	
	泊松假设	正态假设	伽玛假设	逆高期假设	泊松、伽玛	正态、逆高斯
Intercept	0.0347	0.0322	5181.7598	5183.3145	179.5942	166.7841
K_1	0.5617	0.5822	0.9722	0.9713	0.5461	0.5655
K_2	0.6950	0.7294	1.0000	1.0000	0.6950	0.7294
K_3	0.7741	0.7810	1.0000	1.0000	0.7741	0.7810
K_4	0.8424	0.8289	1.0000	1.0000	0.8424	0.8289
K_5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
B_1	3.7701	3.8129	0.8813	0.8816	3.3224	3.3615
B_2	2.3352	2.3639	0.9237	0.9231	2.1570	2.1821
B_3	1.8851	1.9157	0.9477	0.9471	1.7866	1.8143
B_4	1.6484	1.6616	0.9339	0.9344	1.5394	1.5526
B_5	1.4942	1.5067	0.9172	0.9178	1.3705	1.3828
B_6	1.3964	1.4099	0.9474	0.9469	1.3230	1.3350
B_7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Z_1	2.0772	2.1856	1.0000	1.0000	2.0772	2.1856
Z_2	1.6371	1.6889	1.0000	1.0000	1.6371	1.6889
Z_3	1.4116	1.4604	1.0000	1.0000	1.4116	1.4604
Z_4	1.1610	1.1943	1.0000	1.0000	1.1610	1.1943
Z_5	1.4995	1.5197	1.0000	1.0000	1.4995	1.5197
Z_6	1.2276	1.2531	1.0000	1.0000	1.2276	1.2531
Z_7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
M_1	1.0697	1.1090	1.0614	1.0606	1.1354	1.1762
M_2	1.1544	1.1849	1.0000	1.0000	1.1544	1.1849
M_3	0.8352	0.8701	1.1459	1.1467	0.9570	0.9977
M_4	0.5566	0.5533	0.8939	0.8960	0.4975	0.4957
M_5	1.2489	1.2991	1.0000	1.0000	1.2489	1.2991
M_6	0.7648	0.7416	1.0000	1.0000	0.7648	0.7416
M_7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
M_8	1.0000	1.0000	1.3335	1.3324	1.3335	1.3324
M_9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

表 2 关于索赔次数的泊松假设和正态假设的拟合优度比较

泊松假设	正态假设				
Criterion	DF	Value	Value/DF	DF	Value	Value/DF
Deviance	2159	2967.0330	0.1003	2159	216.5473	1.3743
Scaled Deviance	2159	2967.0330	1.3743	2159	2182.0005	1.0107
Pearson Chi-Square	2159	3012.9059	0.1003	2159	216.5473	1.3955
Scaled Pearson χ^2	2159	3012.9059	1.3955	2159	2182.0005	1.0107
Log Likelihood		-442484			4325.9344	

表 3 关于次均赔款的伽玛假设和逆高斯假设的拟合度比较

	伽玛假设			逆高斯假设		
Criterion	DF	Value	Value/DF	DF	Value	Value/DF
Deviance	1785	4886.7232	2.7377	1785	1.1631	0.0007
Scaled Deviance	1785	1957.9469	1.0969	1785	1797	1.0067
Pearson Chi-Square	1785	5539.7647	3.1035	1785	1.1234	0.0006
Scaled Pearson χ^2	1785	2219.5988	1.2435	1785	1735.5674	0.9723
Log Likelihood		-16338			-16214	

4 应用广义线性模型应注意的几个问题

4.1 对数据的要求

应用广义线性模型需要充足的经验数据。一般而言,对于个人保险业务,10万以上的风险单位数才算是充足的。此外,最好有两到三年的经验数据,而不是一年的数据,因为基于一年的数据建模容易受到异常事件的影响。把不同地区的数结合在一起建模也会改进模型的稳定性。但是,当一个地区的数据相当充足时,应该为该地区独立建模。如果数据的获取没有问题,对不同类型的损失经验数据应该分别建模。譬如,对汽车第三者责任保险数据和失窃数据分别建模,可以更加清晰地判定各种保险事故的真正影响因素。在许多实际问题中,最后可能需要建立一个综合性模型,但在前期的分析中分别建模也是非常必要的。

4.2 变量的离散化

变量的离散化是指将连续型变量根据其取值分解成离散型变量的过程。虽然连续型变量本身并不需要人为分解,但直接使用连续型变量建模可能会平滑掉经验数据中的一些重要影响因素。因此在建模之前,有必要对连续型变量进行离散化处理,形成相应的分组变量。一般而言,在建模初期,可以将所有变量转化为分组变量使用,但在把连续型变量离散化的过程中,分组的区间应该比较小,同时应保证每个区间包含足够多的数据。譬如将年龄这个连续型变量进行分解时,如果每个年龄都有足够多的保单持有人,则可以把每个年龄作为一组。但实际上,通常需要把若干年龄合并在一起,才能保证每个年龄组都有足够数量的保单持有人。

4.3 交互作用

交互作用是指一个自变量对因变量的影响作用随着另一个自变量的取值而变化。譬如在表 4中,对于车型 1而言,地区 A和地区 B的赔付率之比是 3:2,而对于车型 2而言,地区 A和地区 B的赔付率之比是 1:1。对于地区 A而言,车型 1和车型 2的赔付率之比为 3:1,而对 B而言,加型 1和车型 2的赔付率之比为 2:1,由此可见,在该例中,不可能用一个参数来描述车型 1和车型 2之间的风险差异,还必须考虑地区的影响;也不可能用一个参数来描述地区 A和地区 B之间的风险差异,还必须考虑车型的影响。这就是所谓的交互影响。

表 4 自变量之间的交互作用

	车型 1		车型 2	
	风险单位数	赔付率	风险单位数	赔付率
地区 A	2000	60%	4000	20%
地区 B	1000	40%	2000	20%

如果某些自变量之间存在交互影响,可以在广义线性模型中引入联合变量加以解决。譬

如,在上例中,可以在广义线性模型中引入“车型 *地区”联合变量,该变量将包含 4个水平。联合变量的引入会迅速增加模型的参数。在本例中,由于车型和地区只有两个水平,所以联合变量“车型 *地区”的引入使得模型的参数增加了 3个,如果这两个变量各有 8个水平,则引入联合变量将使模型的参数增加 63个。因此,联合变量的引入一定要慎重,以免增加不必要的模型参数。

[参考文献]

[1] Duncan Anderson etc. A Practitioner's Guide to Generalized Linear Models[J], CAS 2004 Discussion Papers.

[2] Stephen Mildenhall. A Systematic Relationship Between Minimum Bias Procedure and Generalized Linear Models. 1999 Proceedings of the Casualty Actuarial Society.

[3] R.卡尔斯,M.胡法兹,J.达呐, M.狄尼特著.唐启鹤,胡太忠,成世学译,现代精算风险理论[M].北京,科学出版社,2005.3。

[4] 高惠璇,SAS/STAT软件使用手册[M].北京:中国统计出版社. 1997.9。