在 R 中对保险数据建立广义线性模型

摘要： 本文首先简单分析了传统定价方法的局限性，之后介绍了广义线性模型的理论结构。最后运用 R 软件，详细分析了一家欧洲保险公司 1994—1998 年的车险索赔数据，得到了估计的费率结构。而估计的费率结构与实际的费率结构差异较大，说明原来的费率结构已经落后。

关键词：非寿险定价；广义线性模型；R 软件

中图分类号：O212.1；F840.6　　文献标识码：A　　文章编号：

Model Generalized Linear Models for Insurance Data in R

Abstract: We analyze the limitations of traditional pricing methods briefly, then introduce the theoretical structure of generalized linear models. Finally, we use R software to analyze auto insurance claims data from1994 to1998 of a European insurance company in detail, and we get the estimated rate structure. The big difference between the estimated rate structure and the actual rate structure implies that the original rate structure has lagged behind.

Key Words: Non-life Insurance Pricing; Generalized Linear Models; R Software

引言

广义线性模型（GLM）由 Nelder 和 Wedderburn 于 1972 年提出来[1]，经过近二十年的研究 McCullagh 和 Neld 于 1989 年写了一本系统介绍 GLM 的专著[2]。McCullagh 和 Nelder 介绍了指数分布族，GLM 的参数估计方法，假设检验问题以及一些应用实例。

20 世纪 90 年代英国的精算师，把 GLM 引入到非寿险定价中来，此后的 20 多年 GLM 在很多国家的保险定价实践中得到很大发展，大量的专著和文献都在讨论这个问题。Ohlsson 和 Johansson 介绍了 GLM 在定价中的应用，但又不局限于此，还给出了一些基于 GLM 的扩展模型在非寿险定价中的应用[3]。de Jong 和 Heller 介绍了如何用 GLM 分析保险数据，还给出了大量的实例，进行了细致的分析，是 GLM 在定价中的经典之作[4]。书中是用在保险公司比较流行的 SAS 软件完成数值计算的，但从研究的角度讲 R 软件是一个非常好的选择，而且这方面的书籍文献也很多。Faraway 通过实例用 R 完成了 GLM 中常见模型的估计和假设检验问题[5]。

2010 年，保监会出台了《关于在深圳开展商业车险定价机制改革试点的通知》，明确规定“各财产保险公司可使用现行的商业车险行业指导条款和费率，也可自主开发基于不同客户群体、不同销售渠道的商业车险深圳专用产品”，这为广义线性模型在中国车险定价中的应用提供了制度上的保障。

一、传统的非寿险定价方法

（一）单项分析法

单项分析法是传统的非寿险定价方法，它通过分析特定费率因子的相对赔付率或相对纯保费来确定费率结构。只有当各个费率因子相互独立时，单项分析法才能得出可靠的结论，否则风险分布不均匀可能导致定价结果的严重扭曲，Holler、Sommer 和 Trahair 给出了一个经典例子[6]。

单项分析法仅考虑一些简单的比率关系，其最大缺点是无法提供一个完整的统计分析框架。在非寿险市场竞争日益激烈的情况下，这种定价方法显得过于简单。

（二）最小偏差法

最小偏差法首先确定一个需要最小化的偏差函数，据此求解各个费率因子的相对费率。

最小二乘法、最小 χ^2 法、边际总和法和直接法都是常见的最小偏差法。在仅有两个费率因子的情况下，这些方法都可以写成如下的加权边际总和法：

$$\begin{cases} \sum_i e_{ij} w_{ij} \left(\alpha_i \beta_j - y_{ij} \right) = 0 \\ \sum_j e_{ij} w_{ij} \left(\alpha_i \beta_j - y_{ij} \right) = 0 \end{cases}$$

其中 e_{ij} 为费率单元 (i, j) 的风险单位数，w_{ij} 为权重，α_i、β_j 分别为相应费率因子不同水平的相对费率，y_{ij} 为观测值。在分析损失频率、损失强度等不同情景下，e_{ij}，y_{ij} 代表的意义不同。

相对于单项分析法，最小偏差法解决了风险分布不均匀可能导致的扭曲，但是依然无法提供一个完整的统计分析框架。直到 1992 年，Brockman 和 Wright 首次系统地将广义线性模型引入到非寿险精算定价中来[7]。之后由于一些国家宽松的监管环境和非寿险费率的逐步市场化，广义线性模型定价法得到更多的研究和应用并逐步成为行业标准。实际上，最小偏差法可看做广义线性模型的特例。1999 年，Mildenhall 详细分析了最小偏差法和广义线性模型之间的关系[8]。

二、广义线性模型（GLM）的理论结构

（一）指数型分布（EDF）

概率函数具有如下表达式的分布，称为指数型分布：

$$f_{Y_i}\left(y_i;\theta_i,\phi\right) = \exp\left\{ \frac{y_i\theta_i - b\left(\theta_i\right)}{\phi / w_i} + c\left(y_i,\phi,w_i\right) \right\} \qquad \text{式（1）}$$

其中，θ_i、ϕ 分别称为自然参数、分散参数。$b(\theta)$ 称之为累积量函数。选定累积量函数之后，该指数型分布就是一般的参数化分布，且参数就是自然参数和分散参数。在非寿险定价中，指数型分布往往用其均值和方差来确定。从 2.2 节中的分析可知，自然参数、分散参数通过累积量函数与分布的均值、方差之间存在着一一对应关系。对不同的观测值，自然参数可以不同，但通常假定分散参数和累积量函数相同。当使用极大似然法估计 GLM 时，ϕ 的取值并不影响模型参数的估计，从这个角度讲我们并不关心 ϕ 的取值。如果要作假设检验，就必须估计 ϕ，在 3.2 节中会讨论 ϕ 的估计方法。

（二）指数型分布的均值、方差

推导指数型分布的均值、方差有多种不同的方法，此处使用矩母函数法。设Y_i是指数型分布，具有密度函数式（1），则其矩母函数定义为：

$$M_{Y_i}(t)=\mathrm{E}\left[e^{tY_i}\right]$$

$$=\int e^{ty_i} f_{Y_i}(y_i;\theta_i,\phi)\,dy_i$$

$$=\int \exp\left(\frac{y_i t+\frac{y_i\theta_i-b(\theta_i)}{\phi/w_i}+c(y_i,\phi,w_i)}{}\right)dy_i$$

$$=\int \exp\left(\frac{\frac{y_i(\theta_i+t\phi/w_i)-b(\theta_i+t\phi/w_i)}{\phi/w_i}+c(y_i,\phi,w_i)}{}\right)dy_i$$

$$\times\exp\left(\frac{b(\theta_i+t\phi/w_i)-b(\theta)}{\phi/w_i}\right)$$

上式最后一部分的积分项是指数型分布密度的积分，其值为 1，因此得到指数型分布的矩母函数为：

$$M_{Y_i}(t)=\exp\left(\frac{b(\theta_i+t\phi/w_i)-b(\theta)}{\phi/w_i}\right)$$

累积量函数定义为矩母函数的自然对数：$\psi_{Y_i}(t)=\ln(M_{Y_i}(t))$。累积量函数的一阶、二阶导数在 0 处的取值分别对应分布的均值、方差。这样得到指数型分布的均值、方差公式为：

$\mathrm{E}[Y_i]=\psi'_{Y_i}(0)=b'(\theta_i)$，$\mathrm{Var}[Y_i]=\psi''_{Y_i}(0)=\frac{\phi}{w_i}b''(\theta_i)$。故有：

$$\theta_i=b'^{-1}(\mu_i),\phi=w_i\frac{\mathrm{Var}[Y_i]}{b''\left(b'^{-1}(\mu_i)\right)}$$

常见的分布，如正态分布、伽玛分布、逆高斯分布、泊松分布、二项分布等均是指数型分布。根据这些常见分布，可以通过两种基本方式构造不同的指数型分布，即对原随机变量乘以一个常数α 或使用 Esscher 变换。

假设Y_i是指数型分布，且参数为(θ_i,ϕ)，则$Y_i^*=\alpha Y_i$也是指数型分布，且与随机变量Y_i具有相同的累积量函数，参数为$(\theta_i,\alpha\phi)$。如果对Y_i作 Esscher 变换，且 Esscher 变换参数为h，则新的随机变量也是指数型分布，且具有相同的累积量函数，参数为$(\theta_i+h\phi,\phi)$。Rob

Kaas、Marc Goovaerts 等人详细地分析过这两种方法[9]。Jörgensen 更加细致地讨论了指数型分布的有关性质[10]。

（三）联结函数

在经典线性回归模型中，响应变量的均值与解释变量的线性组合之间存在简单的相等关系，即 $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$。对于 GLM，假设 $\mathbf{g(\mu)} = \mathbf{X\beta}$，$\mathrm{E}[\mathbf{Y}]=\boldsymbol{\mu}$，其中单调函数 $g(x)$ 称之为联结函数。

在分析保险数据时，经常采用对数函数作为联结函数，即 $g(y) = \ln y$。这可以保证响应变量均值的预测值为正，同时乘法模型在很多方面更加适合于保险数据。Brockman 和 Wright 分析了乘法模型的优越之处[7]。

另一个常用的联结函数是 Logit 函数，即 $g(y) = \ln \dfrac{y}{1-y}$。在分析续保率和新业务转换率时，常用 Logit 联结函数。Logit 联结函数最早应用于生物医学统计分析。Logit 联结函数将区间 $(0,1)$ 映射成 $(-\infty,+\infty)$，适用于取值在区间 $(0,1)$ 的响应变量。

概率联结函数，即 $g(y) = \Phi^{-1}(y)$，其中 $\Phi^{-1}(y)$ 为标准正态分布的分布函数的反函数。与 Logit 联结函数一样，概率联结函数可确保响应变量均值的预测值在区间 $(0,1)$。在分析保险数据时，很少选用概率联结函数。

如果联结函数满足 $g(y) = b'^{-1}(y)$，则称之为正则联结函数。正则联结函数可简化模型参数的估计过程，3.1 节的分析表明：当选用正则联结函数时，Newton-Raphson 算法和 Fisher 得分算法的计算结果相同。

虽然理论上可以构造多种不同形式的联结函数，但实际中使用的联结函数种类非常有限。当模型的拟合效果很差时，往往首先考虑解释变量的选择是否正确，响应变量的分布是否合适等，而很少检验联结函数是否正确。

（四）方差函数

在 2.2 节中，得到了：

$$\mathrm{Var}[Y_i] = \frac{\phi}{w_i} b''(\theta_i),\quad \mu_i = \mathrm{E}[Y_i] = b'(\theta_i)$$

则$\theta_i = b'^{-1}(\mu_i)$，故$\mathrm{Var}[Y_i] = \frac{\phi}{w_i} b''(b'^{-1}(\mu_i))$。令$\mathrm{Var}[Y_i] = \frac{\phi}{w_i} V(\mu_i)$，则

$$V(\mu_i) = b''(b'^{-1}(\mu_i))$$

上式称为方差函数，它明确表达了指数型分布的均值—方差结构。可见在 GLM 中，不但响应变量的均值与解释变量之间存在关系，方差也与解释变量之间存在直接的关系，这与经典线性回归模型中假定同方差性具有实质上的不同。

当选用正则联结函数时，

$$\begin{aligned}\mathrm{Var}[Y_i] &= \frac{\phi}{w_i} b''\left(b'^{-1}(\mu_i)\right) \\ &= \frac{\phi}{w_i} b''\left(g(\mu_i)\right) \\ &= \frac{\phi}{w_i} b''\left(\sum_j x_{ij}\beta_j\right)\end{aligned}$$

此式明确了响应变量的方差与解释变量之间存在的直接关系。

常见分布的方差函数具有$V(\mu) = \mu^p$的形式，具有这种均值—方差结构的分布称之为 Tweedie 分布类。2010 年，Esbjorn Ohlsson 和 Bjorn Johansson 给出了 p 取不同值时所对应的分布如表 1[3]。早在 1984 年 Tweedie 就详细地分析了 Tweedie 分布类，但是他分析中的一个错误到 1987 年被纠正。

表 1：Tweedie 分布类

p 的取值	分布类型	支集	分布名称
$p < 0$	连续型	实数	—
$p = 0$	连续型	实数	正态分布
$0 < p < 1$	不存在	—	—
$p = 1$	离散型	自然数	泊松分布
$1 < p < 2$	混合型	非负实数	复合泊松分布
$p = 2$	连续型	正实数	伽玛分布

$2 < p < 3$	连续型	正实数	—
$p = 3$	连续型	正实数	逆高斯分布
$p > 3$	连续型	正实数	—

当$0 < p < 1$时，不存在相应的指数型分布。对保险数据一般不选用 $p < 0$ 。因此在 GLM 中，一般选用 $p \geq 1$ 。当 $p = 3$ 时，相应的分布为逆高斯分布。逆高斯分布比伽玛分布具有更严重的右偏性。对数正态分布比逆高斯分布右偏性更加严重，但对数正态分布不属于指数分布族。对火灾、碰撞等风险，对数正态分布对数据的拟合优于伽玛分布、逆高斯分布。2006 年 Zuanetti, C. Diniz 和 J. Leite 用对数正态分布拟合了保险数据，并给出了参数估计方法[11] 。当$1 < p < 2$ 时，相应的分布是复合泊松分布，此时在 GLM 的框架下可以完成模型的估计，而不用分别考虑损失频率和损失强度。

（五）方差函数与指数型分布的关系

给定的指数型分布一定对应某种形式的均值—方差结构，即存在一个方差函数，但并不是所有的均值—方差结构都对应着某种指数型分布。从表 1 中可知，当 $V(\mu) = \mu^p$，$0 < p < 1$ 时，不存在相应的指数型分布。因此指数型分布能包括的分布类型依然有限。即便对于看似简单的幂指数型方差函数，1983 年 Shaul K Bar-lev 和 Peter Enis 分析方差函数与指数型分布之间的关系时也犯了错，他们认为当 $p < 0$ 时不存在对应的指数型分布[12] 。到 1987 年 Jörgensen 纠正了这个问题[10] 。对于具有幂指数型方差函数的指数型分布，1997 年 Jörgensen 给出了系统的分析[13] 。2006 年 Rigby 和 Stasinopoulos 给出的分布类型远远超出了指数型分布，可以进一步研究这些分布是否可以应用到非寿险定价中去[14] 。

（六）GLM

此时可以完整地给出 GLM 的数学表达式：

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j,\ i = 1, 2, \cdots, n,\ y_i \sim \text{EDF}$$

其中 $\beta_1, \cdots, \beta_p$ 是需要估计的参数，$\{y_i,\ i = 1, \cdots, n\}$ 相互独立。

在分析保险索赔数据时，对损失频率可以使用泊松回归，对损失强度可以使用伽玛回归。后面的实例研究采用这种思路。

三、GLM 参数估计

在经典线性回归模型中，通过最小化 $\sum_{i=1}^{n}(y_i-\mu_i)^2$ 求得参数，即所谓的最小二乘法，并且最小二乘法和极大似然法的估计结果相同。在 GLM 中，使用极大似然法估计模型参数，且极大似然法可以表示成最小化下式的“加权距离”：

$$D=\sum_i w_i d\left(y_i,\mu_i\right),\ d\left(y,\mu\right)=2\left[yb'^{-1}\left(y\right)-b\left(b'^{-1}\left(y\right)\right)-yb'^{-1}\left(\mu\right)+b\left(b'^{-1}\left(\mu\right)\right)\right]$$

上式中的 D 称为离差，是衡量模型拟合效果的统计量。当 y 固定时，$d(y,\mu)$ 在 $\mu=y$ 处取得最小值，这符合距离的直观感觉。定义 $D^*=\frac{D}{\phi}$，这是后文估计离散参数 ϕ 用到的一个统计量。Piet de Jong 和 Gillianz Heller[4]，Esbjorn Ohlsson 和 Bjorn Johansson[3] 从不同的角度解释了 D、D^*。

（一）参数 $\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 的估计

只要能够计算出对数似然函数的一阶、二阶导数，就可以使用 Newton-Raphson 算法和 Fisher 得分算法，并且在选用正则联结函数时两种算法的结果一样。对数似然函数定义为：

$\ell(\boldsymbol{\beta},\phi)=\sum_{i=1}^{n}\ln\ f_{Y_i}\left(y_i;\theta_i,\phi\right)$。

注意到各个参数之间的关系，下述的链式法则是推导中最重要的技巧：

$$\frac{d\ell}{d\beta_j}=\sum_{i=1}^{n}\frac{d\ell}{d\theta_i}\frac{d\theta_i}{d\eta_i}\frac{d\eta_i}{d\beta_j}$$

其中，

$$\left(\frac{d\theta_i}{d\eta_i}\right)^{-1}=\frac{d\eta_i}{d\theta_i}=\frac{d\eta_i}{d\mu_i}\frac{d\mu_i}{d\theta_i}=g'\left(\mu_i\right)V\left(\mu_i\right)$$

$$\frac{d\ell}{d\theta_i}=\frac{y_i-b'(\theta_i)}{\phi/w_i}=\frac{y_i-\mu_i}{\phi/w_i}$$

$$\frac{d\eta_i}{d\beta_j}=x_{ij}$$

故

$$\frac{d\ell}{d\beta_j}=\sum_{i=1}^{n}\frac{y_i-\mu_i}{\phi/w_i}\frac{1}{g'(\mu_i)V(\mu_i)}x_{ij}$$

使用同样的方法求二阶导数：

$$\begin{aligned}\frac{d^2\ell}{d\beta_j\beta_k}&=\sum_{i=1}^{n}\frac{w_i}{\phi}\frac{d}{d\mu_i}\left[\frac{y_i-\mu_i}{V(\mu_i)g'(\mu_i)}\right]x_{ij}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_k}\\&=\sum_{i=1}^{n}\frac{w_i}{\phi}\frac{d}{d\mu_i}\left[\frac{y_i-\mu_i}{V(\mu_i)g'(\mu_i)}\right]x_{ij}\frac{1}{g'(\mu_i)}x_{ik}\\&=-\sum_{i=1}^{n}x_{ij}a_ix_{ik}\end{aligned}$$

其中

$$a_i=\frac{w_i}{\phi V(\mu_i)g'(\mu_i)^2}\left(1+(y_i-\mu_i)\frac{\left[V(\mu_i)g''(\mu_i)+V'(\mu_i)g'(\mu_i)\right]}{V(\mu_i)g'(\mu_i)}\right)$$

设$\mathbf{H}$表示对数似然函数的二阶导数矩阵，$\mathbf{A}=\mathrm{diag}(a_i)$为对角阵，则$\mathbf{H}=-\mathbf{X}'\mathbf{A}\mathbf{X}$。则Fisher信息矩阵$\mathbf{I}$为：

$$\mathbf{I}=-\mathrm{E}\left[\mathbf{H}\right]=\mathbf{X}'\mathrm{E}\left[\mathbf{A}\right]\mathbf{X}=\mathbf{X}'\mathbf{D}\mathbf{X}$$

其中，$\mathbf{D}=\mathrm{diag}(d_i)$，$d_i=\frac{w_i}{\phi V(\mu_i)g'(\mu_i)^2}$。

Newton-Raphson算法的递推公式为：

$$\boldsymbol{\beta}^{(n+1)}=\boldsymbol{\beta}^{(n)}-\left(\mathbf{H}^{(n)}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^{(n)}\mathbf{y}-\mathbf{X}'\mathbf{W}^{(n)}\boldsymbol{\mu}^{(n)}\right)$$

其中，

$$\mathbf{W}=\mathrm{diag}\left(w_i^*\right),\ w_i^*=\frac{d\ell}{d\beta_j}=\frac{w_i}{\phi}\frac{1}{g'(\mu_i)V(\mu_i)}$$

通过不断地计算对数似然函数的一阶、二阶导数值，最后得到参数的收敛值。

Fisher得分算法只是将$\mathbf{H}$换成Fisher信息矩阵$\mathbf{I}$，故得到递推公式为：

$$\boldsymbol{\beta}^{(n+1)}=\boldsymbol{\beta}^{(n)}+\left(\mathbf{I}^{(n)}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^{(n)}\mathbf{y}-\mathbf{X}'\mathbf{W}^{(n)}\boldsymbol{\mu}^{(n)}\right)$$

当选用正则联结函数时，$\mu_i=b'(g(\mu_i))$。两端关于μ_i求导得到：

$$1 = b''\left(g\left(\mu_i\right)\right) g'\left(\mu_i\right)$$

$$g'\left(\mu_i\right) = \frac{1}{b''\left(g\left(\mu_i\right)\right)}$$

$$g''\left(\mu_i\right) = \frac{-b'''\left(g\left(\mu_i\right)\right) g'\left(\mu_i\right)}{b''\left(g\left(\mu_i\right)\right)^2}$$

而方差函数为$V(\mu_i) = b''(\theta_i) = b''\left(b'^{-1}\left(\mu_i\right)\right) = b''\left(g\left(\mu_i\right)\right)$，关于$\mu_i$求导得到：

$$V'\left(\mu_i\right) = b'''\left(g\left(\mu_i\right)\right) g'\left(\mu_i\right)$$

则

$$V\left(\mu_i\right) g''\left(\mu_i\right) + V'\left(\mu_i\right) g'\left(\mu_i\right) = 0$$

因此$\mathbf{A} = \mathrm{E}[\mathbf{A}] = \mathbf{D}$，对比 Newton-Raphson 算法的递推公式和 Fisher 得分算法的递推公式可知，当选用正则联结函数时，两种算法的计算结果一样。

在编写计算程序时，需要考虑算法的选择的问题，这里不再详细讨论这一问题。

（二）分散参数ϕ的估计

对于 GLM，Pearson χ^2 统计量为

$$\mathrm{X}^2 = \sum_i \frac{\left(y_i - \hat{\mu}_i\right)^2}{\mathrm{Var}\left(Y_i\right)} = \frac{1}{\phi}\sum_i w_i \frac{\left(y_i - \hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)}$$

且其近似分布为χ^2_{n-p}分布。因此ϕ的一个近似无偏估计量为：

$$\hat{\phi}_{\mathrm{X}} = \frac{\phi \mathrm{X}^2}{n-p} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)^2}{n-p}$$

而$\frac{D}{\phi}$的近似分布为χ^2_{n-p}分布，因此ϕ的又一个近似无偏估计量为：

$$\hat{\phi}_D = \frac{\phi D^*}{n-p} = \frac{D}{n-p}$$

还可以使用极大似然法来估计分散参数ϕ，这种方法依然依赖于数值计算方法。McCullagh 和 Nelder 考虑了这些方法的选择问题[2]。

四、GLM 的其他专题

前面分析了 GLM 最基本的内容，但是对于实际应用有更多的问题需要考虑。没有绝对好的模型，特定的模型只适合特定种类的数据。因此在实际应用中经常需要在不同的模型间选择，通常的选择标准是 AIC 准则和 BIC 准则。关于这些准则的数学表达式，不再详细讨论，在后面的实例研究中会进一步分析。

在 GLM 中，假设检验主要用于嵌套模型的选择。在经典线性回归模型中，这种检验是精确的，而对于指数型分布这种检验是近似的。

参数的置信区间也是值得关注的一个问题。和假设检验一样，在 GLM 中只能得到近似的置信区间。有时这种“近似”会带来扭曲的结果。如果没有更好的办法，这种“近似”至少可供参考。

解释变量的显著性检验，即解释变量的选择问题在后面的实例研究中会具体说明。

五、实例研究

这里使用的数据是欧洲一家保险公司 1994—1998 年的车险数据。在[2] 中给出了变量名的含义，可直接访问 http://www2.math.su.se/~esbj/GLMbook/mccase.txt 得到这些数据。本文对于变量名作了一些改动，使其更加直观。共有 16383 个观测值，9 个变量，其中有些观测值的保单持有期为零，这一部分观测值需要删除。同时本文采用该保险公司 1995 年费率因子水平的划分办法，在建立模型时也仅仅考虑地区（zon）、MC 类别（mcclass）、车龄（vehicleage）、折扣级别（bonusclass）等四个费率因子，见表 2。随着对数据的不断处理这些变量名会带上前缀，但实际含义不变。正如 2.6 节中提到的，用泊松回归分析损失频率，用伽玛回归分析损失强度。

有很多统计软件可以完成本文中的统计分析，比如 SAS、S-plus 等，鉴于 R 软件及时的更新，从理论角度上讲 R 具有一定的优越之处。特别对于精算研究，R 更具有广阔的应用空间。因此这一节主要考虑如何通过 R 来完成 GLM 的统计分析。除了前面提到的[5]，此部分还参考了 Wood[15]。

在文章的附表中，给出了最基本的 R 代码，有助于理解估计的相对费率是如何得到的。但限于篇幅，没有给出所有的代码和输出结果。附表中上面部分的程序在 R 中可以直接运行，得到分析损失频率的模型结果。中间部分和下面部分的代码只是用来说明程序最重要的部分，要在 R 上运行得到结果还必须有其他类似上面部分的代码。

（一）损失频率

假设 $N_i \sim \text{Poisson}(\lambda_i)$，$e_i$ 为费率单元 i 的风险单位数。一般建立比率 $\frac{N_i}{e_i}$ 与解释变量之间的 GLM。在泊松回归下，风险单位数相当于“抵消项（offset）”。故模型的一般形式为：

$$\log(\lambda_i) = \log(e_i) + \mathbf{X}_i\boldsymbol{\beta}$$

其中 $\mathbf{X}_i$ 为矩阵 $\mathbf{X}$ 的第 i 行。

从原始数据集 mccase.txt 到分析损失频率所用的数据集 vmccase.txt 需要综合使用各种循环语句和条件语句。限于篇幅此处不给出代码，只分析最后结果。表 2 给出了费率因子及不同的水平和该保险公司 1995 年的实际费率结构，其中相对费率为 1 的是基础水平（Base Level）。一般选择风险单位数最多的为基础水平。因此选择费率单元（4,3,3,1）为基础水平。基础水平的选择不影响最终费率结构的估计。表 2 的第三列是 $\exp(\hat{\boldsymbol{\beta}})$，其中 $\hat{\boldsymbol{\beta}}$ 是估计得到的模型参数。

地区（zon）的水平 7 由于赔付记录少，因此其频率相对值为 1，同时输出结果也表明该水平不具有显著性。地区 5、6 的损失频率相对值较少，说明可能不具有显著性，而输出结果证实了地区 5、6 确实不具有显著性。MC 类别的水平 1、4 频率相对值较小，同时输出结果也说明不具有显著性，可以考虑合并水平 1 和水平 4。车龄的各个水平均是显著的，因为车龄较大一般行驶的里程较少，故风险较低而具有较小的相对值。

而折扣级别对于损失频率和损失强度的影响是不同的。折扣级别对于损失频率几乎没有影响，但是对损失强度具有一定的影响。这与无赔款优待系统的内在激励机制是一致的。

（二）损失强度

损失强度是在发生赔案条件下的案均赔款，因此需要删除数据集 vmccase.txt 中赔案数目为零的观测值，这样得到分析损失强度的数据集 vvmccase.txt。在伽玛模型中，为便于解释选用对数联结函数。类似于 5.1 节中的分析，损失强度相对值见表 2 的第四列。特别地，由于地区 7 从没有发生过赔款，因此其相对值选定为基础水平的相对值 1。关于解释变量显著性的分析，类似 5.1 节，此处不再赘述。

表 2：模型结果

费率因子	水平	频率相对值	强度相对值	估计相对费率	实际相对费率
	1	4.6278	1.3221	6.118	7.678
	2	2.2005	1.2498	2.750	4.227

zon	3	1.5792	1.0195	1.610	1.336
	4	1.0000	1.0000	1.000	1.000
	5	1.2621	0.8442	1.065	1.734
	6	1.5532	1.1221	1.743	1.402
	7	1.0000	**1.0000**	1.000	1.402
mcclass	1	1.3695	0.8079	1.106	0.625
	2	1.5948	0.7762	1.238	0.769
	3	1.0000	1.0000	1.000	1.000
	4	1.1068	1.2563	1.390	1.406
	5	1.6716	1.2598	2.106	1.875
	6	2.5586	1.5765	4.034	4.062
	7	2.6180	1.8044	4.724	6.873
vehicleage	1	3.3843	3.1585	10.689	2.000
	2	1.9232	2.4233	4.660	1.200
	3	1.0000	1.0000	1.000	1.000
bonusclass	1	1.0000	1.0000	1.000	1.000
	2	1.0692	1.1261	1.204	0.9
	3	0.8992	1.5146	1.362	0.8

（三）图形分析：损失频率

R 软件可以对模型提供多种多样的图形，能够非常直观地反映模型的优劣。在分析损失频率时，响应变量大部分取值为 0，因此直接调用 *plot(model)*给出的图形非常扭曲，见下图。当响应变量取值非常集中时经常遇到这个问题，因此需要对图形作进一步处理。从“Residuals vs Levelage”图中可知，当对损失频率建立泊松模型时，没有异常观测值。对于异常观测值还可以用 *halfnorm()*来检测，图 5 给出了 5.1 节中建立的泊松模型的 Jacknife 残差的半正态图（halfnorm），同样说明在泊松模型下观测值没有出现异常，这是模型有效的证据。

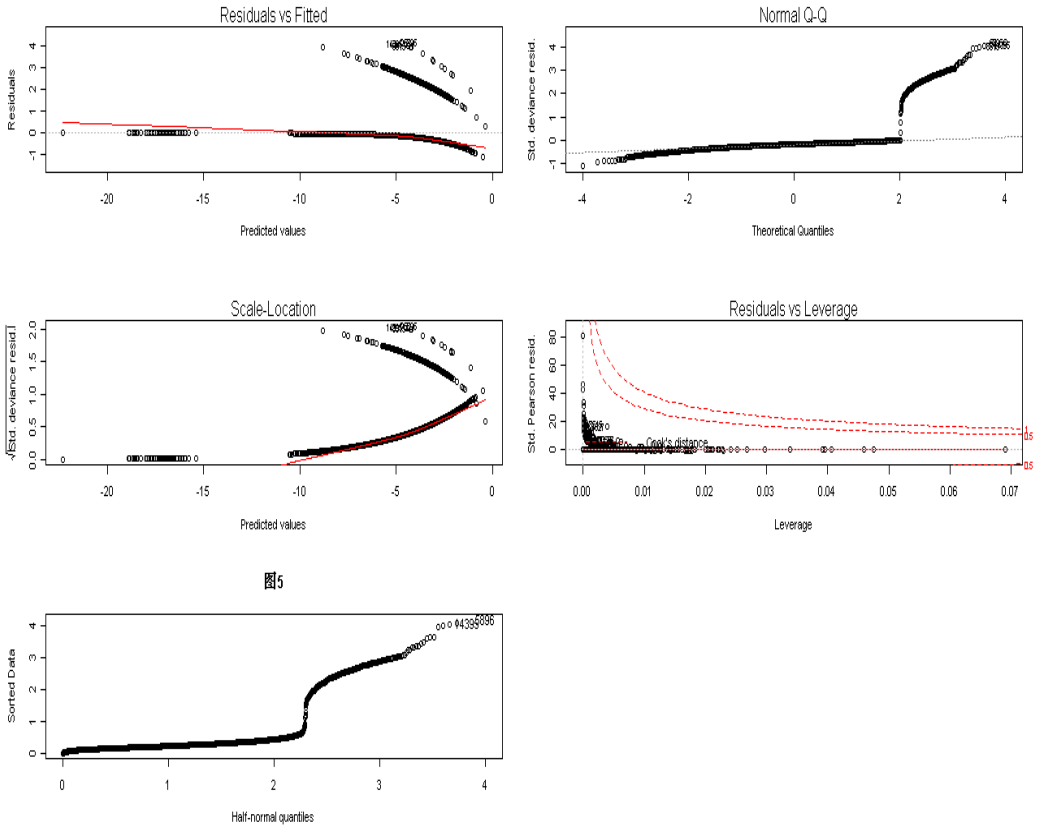


图5

（四）图形分析：损失强度

下图给出了损失强度的有关图形，类似于 5.3 节的分析，这些图形表明对损失强度建立的伽玛模型是合适的。关于这些图形中的一些统计量，比如 Cook 统计量、Levelage 等可参见[4],[15]。由于此例中的费率因子均当成因子来处理，而没有连续型的定价变量，因此虽然可以用 *termplot()*函数画出偏残差图（partial residuals plot），但是并不直观。如果有连续型的定价变量，可以建立更加复杂的模型，但是增加了解释模型的难度。正如[3]中的分析，当有连续型定价变量时可考虑使用样条函数。2008 年，Grgi ć 介绍了样条函数在非寿险定价中的应用[16]。

关于模型的选择，这里简单地考虑损失频率模型的选择问题。负二项分布的方差比均值大，可能更加适合保险数据。在 R 中必须先加载 MASS 包才能完成负二项分布模型的估计。结果表明负二项分布模型的 AIC 为 3428，仅略微小于泊松模型的 AIC 3440。从实际应用来讲这种改进太微小，故选择比较好解释的泊松模型。有时两个模型的 AIC 相差甚远是由于计算时略去了一些常数导致的，因此不要滥用 AIC 准则。

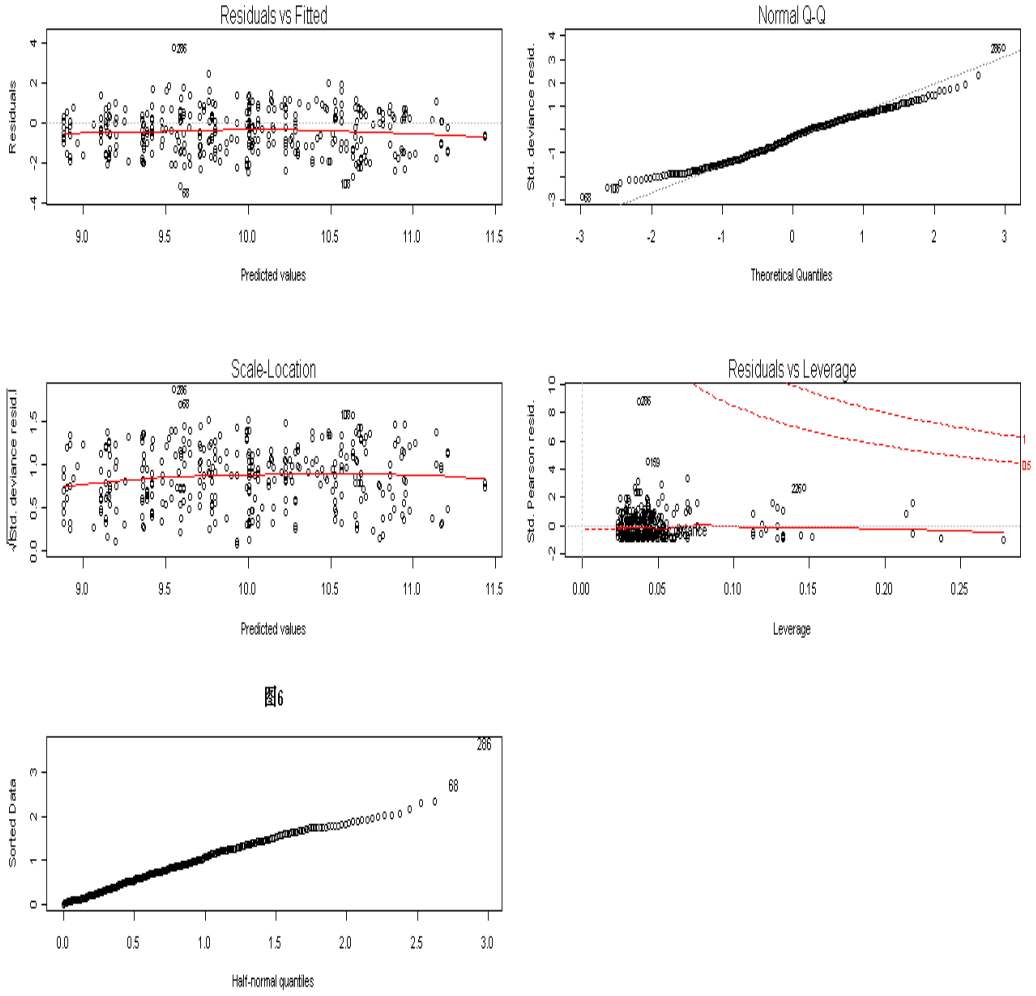
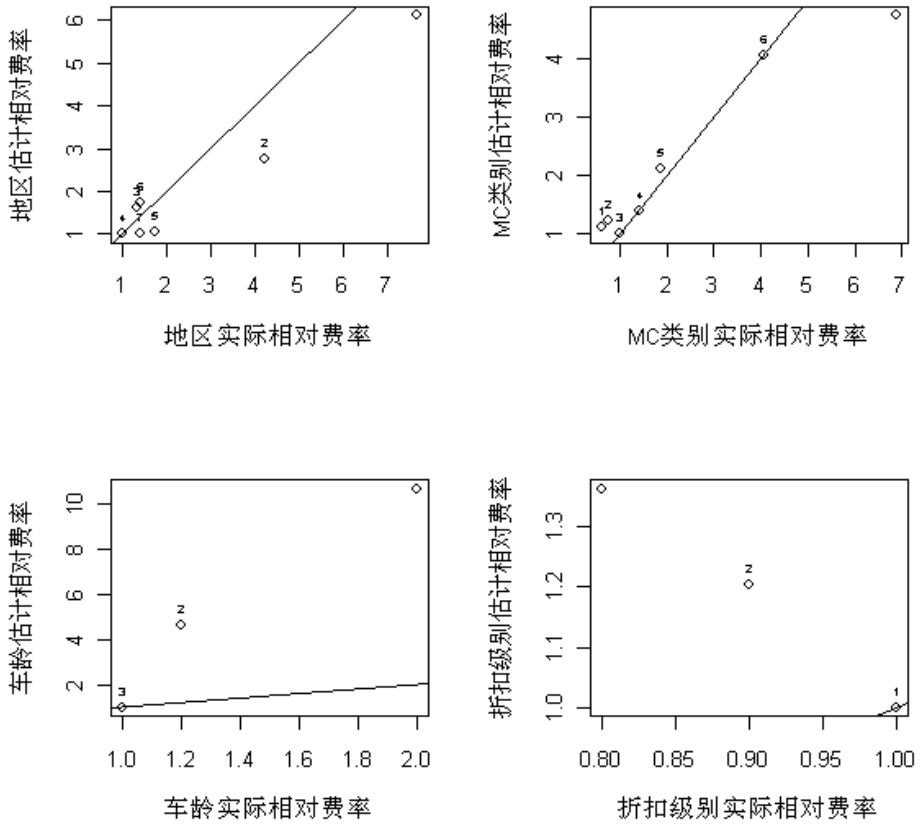


图6

（五）估计的费率结构与 1995 年实际的费率结构

将损失频率相对值与损失强度相对值相乘就得到估计的费率结构，即表 2 的第五列。如何由费率结构得到最终的保险费，需要考虑公司战略、市场定位、外部监管环境、公司利润目标等一系列因素。下面的图形对比了相对费率的估计值和该保险公司 1995 年的实际费率结构（表 2 的第六列）。从图形中可知，车龄和折扣级别的实际相对费率明显低于估计相对费率，这说明可能需要调整这两个费率因子的相对费率结构。限于篇幅，不再考虑因子间的相互作用和其它可能影响费率结构的解释变量，但这些问题对实际定价过程非常重要。



六、结论

本文从不同的角度对 GLM 的理论结构作了细致的分析，某些分析角度确实不同于其他作者。正如正文中提到的，GLM 在将来将成为精算界的标准分析工具，特别是保监会 2010 年出台的《关于在深圳开展商业车险定价机制改革试点的通知》更是为 GLM 在中国保险行业的应用提供了制度上的保障。本文最大的特点是使用了 R 软件的画图功能来形象、直观地给出模型的结果，今后在开发精算软件时可以借用 R 软件中的相关函数。GLM 发展到现在已有很多扩展模型了，最一般的是 GAMLSS，在 www.gamlss.org 上可以找到关于 GAMLSS 的最新研究成果，这也是今后值得研究的一个方向。

附表

```
library(MASS)
library(faraway)
vmccase<-read.table("vmccase.txt",head=T)
attach(vmccase)
M1<-glm(claimsnumber~C(zon,base=4)+C(mcclass,base=3)+C(vvehicleage,base=3)
        +C(bbonusclass,base=1)+offset(log(duration)),family=poisson(link=log))
summary(M1)
op<-par(mfrow=c(3,2)
plot(M1)
halfnorm(rstudent(M1),main="图 5")
```

```
M2<-glm.nb(claimsnumber~C(zon,base=4)+C(mcclass,base=3)+C(vvehicleage,base=3)
        +C(bbonusclass,base=1)+offset(log(duration)))
```

```
claimscostper<-claimscost/claimsnumber
M3<-glm(claimscostper~C(zon,base=4)+C(mcclass,base=3)+C(vvehicleage,base=3)
        +C(bbonusclass,base=1),family=Gamma(link=log))
```

参考文献

[1] J. A. Nelder, R.W. M. Wedderburn. Generalized Linear Models[J]. Journal of the Royal Statistical Society, Series A, 135(3), 370–384.

[2] P. McCullagh, J. A. Nelder. Generalized Linear Models[M].Second Edition. New York: Chapman & Hall, 1989.

[3] Esbjorn Ohlsson, Bjorn Johansson. Non-life Insurance Pricing with Generalized Linear Models[M]. Springer,2010.

[4] Piet de Jong, Gillianz. Heller.Generalized Linear Models for Insurance Data[M]. Cambridge University Press, 2008.

[5] Julian J. Faraway. Extending the Linear Model with R[J].Chapman & Hall/CRC, 2006.

[6] K. D. Holler, D. Sommer, G. Trahair, Something Old, Something New in Classification Ratemaking with a Novel Use of GLMs for Credit Insurance[J]. Casualty Actuarial Forum, Winter 1999

[7] M. J. Brockman, T. S. Wright, Statistical Motor Rating: Making Effective Use of Your Data[J]. JIA , 119 III.

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

[9] Rob Kaas, Marc Goovaerts, Jan Dhaene, Michel Denuit. Modern Actuarial Risk Theory Using R[M].Second Edition. Springer,2008.

[10] B. Jörgensen. Exponential Dispertion Models[J]. J. R. Stat. Soc. Ser. B 49, 1987.

[11] D. Zuanetti, C. Diniz, J. Leite. A Lognormal Model for Insurance Claims Data[J], REVSTAT-Statistical Journal 4(2), 2006.

[12] K. Shaul, Bar-lev, Peter Enis. Reproducibility and Natural Exponential Families with Power Variance Functions[J]. The Annals of Statistics, Vol. 14 No. 4, 1986.

[13] B. Jörgensen. The Theory of Dispersion Models[M]. Chapman & Hall,1997.

14] R. A. Rigby, D. M. Stasinopoulos. A Flexible Regression Approach Using GAMLSS in R[M]. Lecture Notes in University of Lancaster, 2006.

[15] S. N. Wood. Generalized Additive Models: An Introduction with R[M]. Chapman & Hall/CRC, 2006.

[16] V. Grgi ć. Smoothing Splines in Non-life Insurance Pricing[D]. Institute of Mathematical Statistics, Stockholm University, 2008.