汽车保险的精算统计模型

孟生旺 著

中国统计出版社
China Statistics Press

汽车保险的精算统计模型

孟生旺 著

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

汽车保险的精算统计模型 / 孟生旺著. —— 北京 : 中国统计出版社，2014.8

ISBN 978-7-5037-7172-9

Ⅰ. ①汽… Ⅱ. ①孟… Ⅲ. ①汽车保险－统计模型－中国 Ⅳ. ①F842.63

中国版本图书馆 CIP 数据核字(2014)第 171885 号

汽车保险的精算统计模型

作　　者/孟生旺
责任编辑/陈悟朝　姜　洋
封面设计/张　冰
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编号/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://csp.stats.gov.cn
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710mm×1000mm　1/16
字　　数/240 千字
印　　张/15.25
版　　别/2014 年 8 月第 1 版
版　　次/2014 年 8 月第 1 次印刷
定　　价/32.00 元

前言

在非寿险定价的理论和应用研究中，汽车保险的定价问题最受关注，这与财产保险公司的业务结构有关。以我国为例，财产保险公司的保费收入大约60%—70%来自车险业务。

在汽车保险定价模型的研究中，广义线性模型、信度模型、线性混合模型、广义线性混合模型以及它们的各种推广模型都有重要应用。本书基于这些精算与统计模型，重点研究了汽车保险中普通因子和多水平因子的定价问题。

普通因子是指水平数较少的因子，可以用固定效应模型，如广义线性模型或其推广模型进行估计。多水平因子是指水平数很多的因子，如汽车保险中的车系因子和车型因子等，有高达数百甚至数千个水平，建立固定效应模型往往会遇到各种困难。本书讨论了多水平因子定价的三类方法，即信度模型和线性混合模型，广义线性模型与线性混合模型的迭代算法，以及广义线性混合模型。当然，在这些模型基础上还可以进一步衍生出各种拓展模型。

关于汽车保险的定价模型，可以参考的文献很多。本书也以车险定价问题为主，但研究的重点与现有文献有所不同，主要对普通因子和多水平因子的联合定价方法进行探讨。现有文献没有充分考虑普通因子和多水平因子之间的相互影响，这可能导致定价结果出现偏差。

在车险定价实务中，普通因子的定价通常使用广

义线性模型，多水平因子的定价使用信度模型。常用的信度模型可以通过线性混合模型来实现，所以车险定价中最基本的统计模型是广义线性模型和线性混合模型。广义线性模型推广了线性回归模型的分布假设，而线性混合模型在线性回归模型的基础上引入了随机效应。如果把广义线性模型和线性混合模型的假设结合在一起，就得到了所谓的广义线性混合模型，即不仅把线性回归模型中的分布假设从正态分布推广到了指数分布族，而且在模型中引入了随机效应。因此，从理论上讲，应用广义线性混合模型就可以同时解决普通因子和多水平因子的定价问题，但广义线性混合模型在实际应用中有可能遇到收敛性困难，此时可以考虑使用广义线性模型与线性混合模型的迭代算法，这是本书将要重点研究的主题之一。

迭代算法有其自身的优势。实证研究结果表明，广义线性模型与线性混合模型的各种迭代算法都能收敛。如果使用相同的分布假设，迭代算法与广义线性混合模型的结果非常接近。迭代算法的另一个优点是，可以进一步拓展广义线性模型，如在零膨胀泊松分布、零膨胀负二项分布或零调整逆高斯分布的基础上建立回归模型，从而进一步提高模型对实际数据的拟合效果和预测能力，但在通常的广义线性混合模型中，分布假设只能限定在指数分布族。

迭代算法也有其缺陷。该算法应用广义线性模型或其推广模型估计普通因子，应用线性混合模型估计多水平因子，还不能从总体上对模型的拟合优度进行评价。

在汽车保险定价模型的实证研究中，统计软件扮演着至关重要的角色，其中最常使用的两种统计软件是SAS和R。R软件的优点是小巧灵活，绘图功能强大，有各种程序包可以参考使用，容易对现有模型进行扩展，但运行时需要耗费较多内存。当数据量较大时，应用SAS的优势就比较明显。本书在模拟分析中主要使用R软件，实证分析中主要使用SAS软件。对于分析结果的呈现，尽可能使用图示的方式，力求一目了然。

全书包括8章，各章的主要研究内容如下：

第1章对各种车险定价模型进行了简要评述。从最简单的单变量分析法开始，到边际总和法，再到目前最流行的广义线性模型，本章对各种定价模型的性质和特点进行了概括性总结，最后引出了本书研究的重点内容，即多水平因子的定价问题。

第2章首先对普通因子的定价模型进行了概括和总结，介绍了广义线

性模型的参数估计和检验方法，通过模拟数据分析了车险定价中常用的广义线性模型以及它们之间的关系，然后重点研究了普通因子定价中的几个特殊问题，包括过离散索赔次数模型，零膨胀索赔次数模型，费率约束条件下对广义线性模型的推广，以及贝叶斯广义线性模型在车险定价中的应用。

第3章讨论了多水平因子定价的基本理论，包括信度模型、线性混合模型、广义线性模型与线性混合模型的迭代算法，以及广义线性混合模型等，证明了Bühlmann-Straub信度模型是线性混合模型的特例，提出了估计多层信度模型的分步算法，并将其推广到广义线性模型与线性混合模型的迭代算法，最后对固定效应模型和随机效应模型进行了比较研究。这部分以理论分析和模拟研究为主。

第4章应用汽车保险的实际数据，首先在不考虑普通因子影响的情况下，讨论了信度模型与线性混合模型在估计多水平因子中的应用及其相互关系，然后在线性混合模型的框架下，研究了普通因子与多水平因子的相互影响。

第5章研究了车系与车型联合索赔强度因子的估计方法，内容涉及线性混合模型、伽马广义线性模型与线性混合模型的迭代算法、伽马广义线性混合模型，以及索赔强度对数的线性混合模型等，最后对各种模型的结果进行了比较。

第6章研究了车系与车型联合索赔频率因子的估计方法，内容包括：(1)泊松回归与线性混合模型的迭代算法，负二项回归与线性混合模型的迭代算法，零膨胀泊松回归与线性混合模型的迭代算法以及零膨胀负二项回归与线性混合模型的迭代算法；(2)泊松广义线性混合模型和负二项广义线性混合模型；(3)随机效应零膨胀泊松回归模型和随机效应零膨胀负二项回归模型。

第7章研究了车系与车型联合纯保费因子的估计方法，具体内容包括：(1)基于车系与车型的联合索赔频率因子和联合索赔强度因子计算联合纯保费因子；(2)Tweedie回归模型与线性混合模型的迭代算法；(3)随机效应Tweedie回归模型和随机效应零调整逆高斯回归模型在多水平纯保费因子定价中的应用。

第8章对我国汽车保险中的一类特殊业务(即交强险业务)进行了实证分析，主要基于2009—2011年交强险业务的实际数据，分析了交强险保费水平在不同业务类型和不同地区之间存在的公平性问题，揭示了交强险费

目录

第 1 章　汽车保险的定价模型综述　1

1.1　单变量分析法　2

1.2　边际总和法　4

1.3　广义线性模型　6

1.4　广义线性模型的推广　7

1.5　信度模型　8

1.6　多水平因子的定价模型　10

1.7　相依风险的定价模型　11

第 2 章　普通因子的定价　12

2.1　边际总和法　12

2.2　广义线性模型　15

2.2.1　指数分布族　15

2.2.2　广义线性模型的参数估计　17

2.2.3　广义线性模型的评价和检验　20

2.2.4　预测结果的平衡性　22

2.2.5　等价的广义线性模型　24

2.2.6　数据压缩　25

2.2.7　泊松回归与索赔频率预测　26

2.2.8　伽马回归与索赔强度预测　33

2.2.9　Tweedie 回归与纯保费预测　39

2.2.10　Logistic 回归与出险概率预测　44

2.3　过离散索赔次数模型　46

2.3.1　负二项回归　47

2.3.2　泊松－逆高斯回归　48

2.3.3　泊松－对数正态回归　48

2.3.4　广义泊松回归　48

2.3.5　混合负二项回归　49

2.3.6　应用案例　49

2.4　零膨胀和零调整索赔次数模型　52

2.4.1　零膨胀索赔次数模型　52

目录

2.4.2 零调整索赔次数模型 56
2.5 市场约束条件下的车险定价模型 58
2.5.1 等式约束 59
2.5.2 一般线性约束 62
2.5.3 应用案例 64
2.6 考虑先验信息的车险定价模型 68
第3章 多水平因子的定价 74
3.1 信度模型 74
3.1.1 有限波动信度模型 75
3.1.2 Bühlmann 信度模型 79
3.1.3 Bühlmann-Straub 信度模型 81
3.1.4 信度模型的另一种解释 83
3.1.5 信度模型的特例:奖惩系统 84
3.2 信度保费的计算 87
3.3 线性混合模型 91
3.3.1 线性混合模型的一般形式 91
3.3.2 线性混合模型与信度模型的关系 94
3.4 基于线性混合模型的信度保费 96
3.5 多层信度模型 100
3.6 多层信度模型的分步计算 107
3.7 GLM 与信度模型的迭代算法 113
3.7.1 只有一个多水平因子 113
3.7.2 嵌套的多水平因子(方法 I) 114
3.7.3 嵌套的多水平因子(方法 II) 114
3.8 广义线性混合模型 115
3.9 随机效应模型与固定效应模型的比较 118
第4章 普通因子与多水平因子的关系 122
4.1 索赔强度的车系因子 123
4.2 索赔频率的车系因子 137
4.3 索赔强度的车型因子 141
4.4 索赔频率的车型因子 143

目录

4.5 普通因子对多水平因子的影响 145
4.5.1 索赔强度模型 145
4.5.2 索赔频率模型 149
第5章 索赔强度的多水平因子 152
5.1 索赔强度的线性混合模型 152
5.2 伽马回归与 LMM 的迭代算法 155
5.3 伽马广义线性混合模型 159
5.4 索赔强度对数的线性混合模型 163
5.5 模型比较 166
第6章 索赔频率的多水平因子 170
6.1 索赔频率 GLM 与 LMM 的迭代算法 170
6.1.1 泊松回归与 LMM 的迭代算法 170
6.1.2 负二项回归与 LMM 的迭代算法 174
6.1.3 零膨胀泊松回归与 LMM 的迭代算法 175
6.1.4 零膨胀负二项回归与 LMM 的迭代算法 177
6.2 索赔频率的广义线性混合模型 178
6.2.1 泊松 GLMM 178
6.2.2 负二项 GLMM 183
6.3 随机效应零膨胀索赔次数模型 184
第7章 纯保费的多水平因子 187
7.1 独立假设下纯保费的多水平因子 187
7.2 Tweedie 回归与 LMM 的迭代算法 190
7.2.1 Tweedie 回归 190
7.2.2 Tweedie 回归与 LMM 的迭代算法 193
7.3 随机效应 Tweedie 回归模型 195
7.4 随机效应零调整逆高斯回归模型 197
第8章 交强险的费率结构分析 204
8.1 保费的公平性分析 205
8.2 费用率的合理性分析 206
8.3 交强险的市场竞争 208
8.3.1 基于业务类型的市场竞争 208

目录

8.3.2 基于地区分布的市场竞争 210
8.3.3 保险公司的风险选择能力指数 212
8.4 赔付率和费用率的影响因素 213
8.4.1 公司赔付率的影响因素 214
8.4.2 公司费用率的影响因素 215
8.4.3 地区赔付率的影响因素 216

总结与展望 220
索引 223
参考文献 227

第1章

汽车保险的定价模型综述

非寿险精算研究的两个主要问题是保险定价和准备金评估，它们的核心内容都是对未来的赔款和费用进行预测。保险定价是对即将签发的新保单在未来的赔款和费用进行预测，而准备金评估是对已经生效的保单在未来发生的赔款和费用进行预测。在非寿险的损失预测中，既有精算模型，如信度模型，也有统计模型，如广义线性模型。信度模型虽然发源于精算领域，但其基本原理与统计模型并无本质区别。有鉴于此，本书把用于保险损失预测的各种模型统称为精算统计模型。

汽车保险的精算与统计模型理应包含定价和准备金评估两个方面的内容，但本书主要研究汽车保险的定价模型，简称车险定价模型。车险定价模型大致可以分为两类：分类费率模型和经验费率模型。

分类费率是一种平均费率，即首先对个体风险进行分类，然后厘定各个风险类别的平均费率。分类费率的应用可以追溯到保险业发展的最初阶段，譬如在早期的海上保险中，保费与船舶的结构有关，而在火灾保险中，保费与房屋的建造类型有关。随着保险业的发展，分类费率模型经历了一个从简单到复杂，从初级到高级的发展过程。按照时间先后顺序排列，分类费率模型包括单变量分析法、边际总和法、广义线性模型以及它们的各种推广模型。

经验费率是基于个体风险自身的损失经验所厘定的费率，通常表现为应用个体风险的损失经验对某种已知的费率进行调整，譬如对分类费率或当前使用的费率进行调整。经验费率模型主要包括信度模型及其在汽车保险中的一种简化形式，即奖惩系统（Bonus-Malus system，BMS）或无赔款优待系统（No-Claim Discount，NCD）。

分类费率模型主要适用于个人汽车保险业务，而经验费率模型主要适用于团体汽车保险业务。对于个人汽车保险业务，每份个体保单的经验数据往往缺

乏足够的可信度。把个体保单根据其风险特征进行分类汇总，每个类别的经验数据就会比较充足，从而可以较为可靠地厘定各个风险类别的平均费率。对于团体汽车保险业务，如果每份保单承保的车辆数很多，经验数据十分充足，就可以直接应用经验费率模型。当然，这种应用范围的划分不是绝对的。在个人汽车保险业务中，可以在分类费率的基础上应用经验费率模型，如奖惩系统或无赔款优待系统。而在团体汽车保险业务中，也可以结合应用分类费率和经验费率，譬如首先根据风险特征对车队进行分类，然后厘定每个风险类别的平均费率，最后再基于车队自身的损失经验对分类费率进行调整。

本章将主要对汽车保险定价中常用的各种分类费率模型和经验费率模型进行简要评述。

1.1 单变量分析法

在分类费率应用的初期阶段，定价的主要方法是单变量分析法(one-way analysis)，也称作单项分析法。分类变量往往不止一个，但单变量分析法每次仅分析一个变量对损失的影响，不考虑这些变量之间的相互影响。单变量分析法的优点是直观简单，缺陷是当个体风险在各个类别的分布不均匀时可能得出完全错误的结论。

为了说明单变量分析法的基本原理及其特点，假设在汽车保险中仅使用两个分类变量，即汽车的行驶区域和汽车的用途。汽车的行驶区域有两个水平：区域 A 和区域 B。汽车的用途也有 2 个水平：私人用车和商业用车。

在区域 A，私人用车有 1 万个车年数，平均每个车年的损失为 1000 元；商业用车有 1 万个车年数，平均每个车年的损失是 2000 元。

在区域 B，私人用车有 1 万个车年数，平均每个车年的损失是 900 元；商业用车有 1 万个车年数，平均每个车年的损失是 1800 元。

容易看出，不同分类变量的真实风险水平如下：

从汽车的行驶区域来看，无论是私人用车还是家庭用车，区域 B 平均每个车年的损失是区域 A 的 0.9 倍。

从汽车的用途来看，无论是在区域 A 还是在区域 B，商业用车平均每个车年的损失是私人用车的 2 倍。

由此可见，合理的定价结果应该是：区域 B 的费率是区域 A 的 0.9 倍，商业用车的费率是私人用车的 2 倍。

本例数据的一个特点是，两个分类变量把所有汽车划分为 4 个类别，每个

类别都有1万个车年，即个体风险在每个类别中的分布是均衡的。下面应用单变量分析法来厘定每个风险类别的费率，假设基准类别为行驶区域A的私人用车。

首先分析汽车行驶区域变量。在区域A，平均每个车年的损失是(1000+2000)/2 = 1500元；在区域B，平均每个车年的损失是(900+1800)/2=1350元，区域B平均每个车年的损失是区域A的0.9倍，即区域B的费率因子为0.9。

再分析汽车用途变量。对于私人用车，平均每个车年的损失是(1000+900)/2 = 850元；对于商业用车，平均每个车年的损失是(2000+1800)/2=1900元。商业用车平均每个车年的损失是私人用车的2倍，即商业用车的费率因子为2。可见，在该例中，单变量分析法得出的结果与真实情况完全相符。

在前例中，个体风险在每个风险类别的分布是均衡的，所以单变量分析法的结果与实际相符。当个体风险在每个风险类别的分布失衡时，单变量分析法得出的结论就会出现偏差，甚至完全背离实际情况。不妨假设在上例中，区域B的商业用车有2万个车年数(而不是原来假设的1万个车年数)，其他类别仍然只有1万个车年数。此时，区域A平均每个车年的损失为1500元，区域B平均每个车年的损失为(900+1800×2)/(1+2) = 1500元，即区域B与区域A相等，由此求得区域B的费率因子等于1，偏离了真实的费率因子0.9。

表1—1给出了当其他风险类别的车年数保持在10000，而区域B的商业用车的车年数变化时，用单变量分析法计算的区域B的费率因子和商业用车的费率因子，以及区域B的商业用车的纯保费。可见，当区域B的商业用车的车年数大于10000时，单变量分析法会高估区域B的风险，低估商业用车的风险，最终导致高估区域B的商业用车的纯保费；反之，当区域B的商业用车的车年数小于10000时，单变量分析法会低估区域B的风险，高估商业用车的风险，最终导致低估区域B的商业用车的纯保费。

表1—1 车年数变化对费率因子的影响

区域B商业用车的车年数	0	4000	8000	12000	16000	20000
区域B的费率因子(单变量分析法)	0.60	0.77	0.87	0.93	0.97	1
区域B的费率因子(真实值)	0.9	0.9	0.9	0.9	0.9	0.9
商业用车的费率因子(单变量分析法)	2.11	2.05	2.01	1.99	1.98	1.96
商业用车的费率因子(真实值)	2	2	2	2	2	2
区域B商业用车的纯保费(单变量分析法)	1263	1578	1743	1846	1915	1965
区域B商业用车的纯保费(真实值)	1800	1800	1800	1800	1800	1800

由此可见，当个体风险在每个风险类别的分布不均衡时，单变量分析法的结果是不可靠的，有可能偏高，也有可能偏低。偏差的大小与个体风险分布的失衡程度有关，越是失衡的风险分布，单变量分析法产生的偏差会越大。

通过迭代运算可以消除单变量分析法中风险分布不平衡所带来的影响，具体可参见孟生旺(2011)，此处不再赘述。

1.2 边际总和法

当个体风险在各个风险类别中的分布不均衡时，单变量分析法的结果会出现偏差。边际总和法(marginal total method)可以克服不均衡的风险分布所造成的偏差。边际总和法要求通过模型预测的边际损失总和等于实际观察的边际损失总和，因此也称作平衡法(balance method)。边际总和是指根据每个分类变量的不同水平求得的损失之和。譬如在前述的例子中，根据汽车的行驶区域可以分别求得区域A的边际损失总和与区域B的边际损失总和，根据汽车的用途可以分别求得私人用车的边际损失总和与商业用车的边际损失总和。每一个边际损失总和都可以通过两种方法计算，一种是根据实际观察值计算，一种是基于模型的预测值计算。令这两种方法计算的边际损失总和相等，即可建立求解模型参数的方程组。

仍然以汽车保险为例，假设如下：

区域A：私人用车平均每个车年的损失为1000元，商业用车平均每个车年的损失为2000元，私人用车和商业用车各有1万个车年。

区域B：私人用车平均每个车年的损失为900元，商业用车平均每个车年的损失为1800元，私人用车和商业用车分别有1万个车年和k万个车年。这里的k是一个已知值，在下面的分析中根据需要给定。

假设用乘法模型对每个风险类别的损失进行预测，即风险类别(i,j)的损失预测值可以表示为：

$$\mu_{ij}=\mu\alpha_i\beta_j$$

其中，μ表示基准类别的损失预测值，本例假设基准类别为区域A的私人用车，故$\mu=1000$元。α_i是汽车行驶区域因子，下标$i=1$表示区域A，$i=2$表示区域B；β_j是汽车用途因子，$j=1$表示私人用车，$j=2$表示商业用车。

首先根据汽车的行驶区域计算边际损失总和。区域A的边际损失总和可以表示如下：

$$1000\alpha_1\beta_1+1000\alpha_1\beta_2=1000+2000$$

上式左边是基于模型预测的边际损失总和，右边是根据实际观察数据计算的边际损失总和。

类似地，可以求得区域 B 的边际损失总和如下：

$$1000\alpha_2\beta_1 + 1000k\alpha_2\beta_2 = 900 + 1800k$$

私人用车和商业用车的边际损失总和分别为：

$$1000\alpha_1\beta_1 + 1000\alpha_2\beta_1 = 1000 + 900$$

$$1000\alpha_1\beta_2 + 1000k\alpha_2\beta_2 = 2000 + 1800k$$

将上述等式变形以后，即可得到求解汽车行驶区域因子和汽车用途因子的迭代公式如下：

$$\alpha_1 = \frac{1000 + 2000}{1000(\beta_1 + \beta_2)}, \alpha_2 = \frac{900 + 1800k}{1000(\beta_1 + k\beta_2)} \tag{1.1}$$

$$\beta_1 = \frac{1000 + 900}{1000(\alpha_1 + \alpha_2)}, \beta_2 = \frac{2000 + 1800k}{1000(\alpha_1 + k\alpha_2)} \tag{1.2}$$

其中，k 是已知的车年数，需要事先给定。

在应用上述迭代公式时，可以首先给出汽车用途因子 (β_1, β_2) 的初始值，如令 $\beta_1 = 1, \beta_2 = 1$，并由式(1.1)求得汽车行驶区域因子 (α_1, α_2) 的估计值，将其代入式(1.2)，可以求得 (β_1, β_2) 的估计值，再将其代入式(1.1)，可以求得 (α_1, α_2) 的新估计值，如此循环下去，最终会收敛，从而得到汽车行驶区域 B 的费率因子为 α_2/α_1，商业用车的费率因子为 β_2/β_1。

在单变量分析法中，只有当 $k = 1$ 时，即个体风险在各个风险类别的分布是均衡的，才能保证分析结果与实际情况完全相符，即区域 B 平均每个车年的损失与区域 A 平均每个车年的损失之比为 0.9，商业用车平均每个车年的损失与私人用车平均每个车年的损失之比为 2。在风险分布失衡的情况下，即当 $k \neq 1$ 时，单变量分析法的结果都将出现偏差。但在边际总和法中，无论 k 的取值如何变化，经过多次迭代之后，总能求得与实际情况完全相符的结果。

边际总和法是一种迭代算法，与其相类似的方法还有最小卡方法、最小二乘法和直接法等。这类方法可以统称为最小偏差法(minimum bias method)或迭代法。它们优化的目标函数各不相同，因此得到的费率厘定结果也有所差异。譬如，边际总和法厘定的费率总和等于经验损失总和，而最小卡方法厘定的费率总和大于或等于经验损失总和。

最小偏差法虽然可以解决个体风险在不同风险类别中分布失衡所带来的问题，但因为仍然不是完整意义上的统计方法，所以对参数的估计结果不能进行显著性检验，对模型的整体拟合效果也无法进行统计评价。

1.3 广义线性模型

进入21世纪以来，分类费率厘定的主流方法已经被广义线性模型(generalized linear models)所取代。可以证明，泊松分布假设下的广义线性模型与前述的边际总和法得到的费率厘定结果完全相同。但广义线性模型具有边际总和法无可比拟的优势，不仅可以对参数的显著性和模型的整体拟合效果进行统计检验，还可以分析变量的交互效应对潜在损失的影响。

车险定价的基础性工作是对索赔频率、索赔强度或纯保费进行预测。索赔频率是指平均每个车年的索赔次数，索赔强度是指平均每次索赔的赔款金额，纯保费是指平均每个车年的赔款金额，纯保费等于索赔频率与索赔强度的乘积。在应用广义线性模型厘定车险分类费率时，通常假设索赔次数、索赔强度或纯保费服从指数分布族中的某个分布。指数分布族包含了许多保险数据分析中常用的分布类型，如泊松分布、二项分布、伽马分布、逆高斯分布、Tweedie分布等。在预测索赔频率时，通常假设索赔次数服从泊松分布；在预测索赔强度时，通常假设每次的损失金额服从伽马分布或逆高斯分布；在预测纯保费时，通常假设每个车年的损失观察值服从Tweedie分布；而在预测事故发生概率时，通常使用二项分布假设。

广义线性模型的一般形式可以表示如下：

$$E(y_i)=\mu_i=g^{-1}(X_i'\beta)$$

其中，因变量 y_i 表示个体风险 i 的损失观察值，服从指数分布族中的某个分布，其均值为 μ_i。g 是连接函数，g^{-1} 表示其逆函数。X_i' 表示个体风险 i 的解释变量向量，β 是回归系数向量。

在正态分布假设和恒等连接函数下，广义线性模型就退化为普通的线性回归模型。线性回归模型可以通过最小二乘法或极大似然法进行参数估计，两者是等价的。在广义线性模型中，应用极大似然法进行参数估计，但其实现过程等价于迭代加权最小二乘法。

极大似然估计具有渐近正态性，应用广义线性模型不仅可以求得参数估计值，即费率因子，而且可以对每个费率因子的显著性水平进行假设检验，即判定费率因子是否显著不为零。此外，通过偏差(Deviance)、对数似然函数、AIC和BIC等统计量可以对模型的显著性进行检验和比较。

广义线性模型的理论文献很多，最早当属Nelder(1972)的开创性工作，其他可以参考的文献还有Dobson (2008)，Mccullagh (1989)，Venter (2007)，Lee

(2006)等。从本世纪初开始,广义线性模型在非寿险精算实务中得到了广泛应用,譬如可参见 De Jong (2008), Ohlsson (2010), 孟生旺 (2007) 和罗妍 (2011)等。

1.4　广义线性模型的推广

广义线性模型要求因变量服从指数分布族中的某个分布,而实际的保险数据有可能偏离这类分布。一种偏离情况是实际数据的尾部较长,即所谓的厚尾数据,另一种情况是实际数据在零点有较大的概率堆积,即所谓的零膨胀数据。这两种情况下的数据都存在过离散现象,即实际数据的方差大于模型的理论方差。在过离散情况下,广义线性模型对实际数据的拟合效果往往不够理想,还有可能高估参数的显著性。

零膨胀现象在汽车保险的索赔次数数据中比较常见,譬如,某些被保险车辆在保险期间根本没有上路行驶,因此不可能发生任何索赔。此外,由于汽车保险中广泛使用了奖惩系统(MBS),许多车主为了避免索赔造成续保保费的上升,对一些小额的损失事故不向保险公司报案,这也会导致实际的索赔观察数据中存在大量的零值。指数分布族中的分布很难对这些额外的零值进行有效拟合。此时就得考虑零膨胀模型,如零膨胀泊松分布、零膨胀负二项分布、零膨胀泊松—逆高斯分布等。在零膨胀分布假设下拟合实际损失数据的案例研究可参考 Yip (2005)和徐昕 (2012)。

在索赔次数数据中,如果某些保单的索赔次数远远大于平均水平,就出现了所谓的厚尾索赔次数数据。拟合这类索赔次数数据时,合适的分布可能不是泊松分布,而是混合泊松分布,如负二项分布、泊松—逆高斯分布和泊松—对数正态分布,它们的结构函数分别为伽马分布、逆高斯分布和对数正态分布。随着结构函数的尾部变厚,混合泊松分布的尾部也越来越厚。此外,广义泊松分布、混合负二项分布、Delaporte 分布和 Sichel 分布也属于混合泊松分布,适用于拟合厚尾的索赔次数数据。

在损失金额数据的拟合中,当实际数据的尾部很厚时,常用的伽马回归和逆高斯回归也会表现不佳,此时可以考虑使用广义帕累托分布或 GB2 分布。

因为广义线性模型不能很好拟合零膨胀数据和厚尾数据,从而有必要对其进行推广。广义线性模型的推广可以从多个角度展开,譬如,可以推广分布假设,从指数分布族扩展到更加一般的分布类型,也可以在广义线性模型中纳入平滑预测项,将其推广到广义可加模型(generalized additive models, GAM),可