文章编号:1000-5862(2021)04-0353-09

索赔次数的贝叶斯预测与信度近似

章　溢[1],周金亮[2]
(1. 江西师范大学财政金融学院,江西 南昌　330022;2. 江西师范大学数学与统计学院,江西 南昌　330022)

摘要: 该文先对保单的索赔次数建立了贝叶斯模型;然后,根据样本分布和先验分布已知或未知分 3 种情形讨论了索赔次数的贝叶斯预测及信度预测,并给出了结构参数的估计方法;最后,通过数值模拟的方法对 3 个预测的均方误差和收敛性进行了比较.

关键词: 泊松分布;伽马分布;贝叶斯预测;信度近似;结构参数
中图分类号: O 212.9　　**文献标志码:** A　　**DOI:** 10.16357/j.cnki.issn1000-5862.2021.04.06

0　引言

在非寿险精算中,更加精准地预测风险保单的未来索赔次数一直都是精算师需要探讨研究的问题. 在该问题中已知涵盖了多个风险保单的保单组合在过去若干年份内已经发生的索赔次数,目标是基于这些损失记录,能够较好预测保险公司的保单组合在将来可能发生的索赔次数. 假设在保单组合中包含了若干个风险保单,其中第 i 个风险保单的索赔次数依赖于第 i 个风险保单的风险参数 θ_i. 根据风险的非齐次性[1],风险参数 θ_i 被认为是随机变量,并且会服从某种共同的先验分布,记为 $\pi(\theta)$. 本文需要根据保单组合已有的索赔损失的样本数据以及风险参数 θ_i 的先验分布信息,讨论如何更精确地对未来可能发生的保单索赔次数进行预测.

在汽车保险中,保单的索赔次数是设定奖惩系统的依据. 精算师通过分析被保险人在前些年的索赔次数情况,对在未来若干年内可能发生的索赔次数进行预测,进而根据设定一定的保费奖励或惩罚来制定未来的保费折扣,以鼓励更少的索赔发生. 因此,索赔次数的拟合与预测,一直是非寿险精算的热点研究问题. 张连增等[2] 研究了在泊松提升模型中车险索赔频率的预测问题;王选鹤等[3] 构建了零膨胀混合泊松回归模型,并研究了车险索赔次数的预测问题;孟生旺等[4] 研究了在零膨胀索赔次数模型中带有随机效应的回归模型,并给出了未来索赔次数的预测. 其他有关索赔次数的预测在寿险精算中的应用可以参见文献[5-7].

在对保单的未来索赔次数的预测过程中,目标是利用给定的损失函数,在所有样本的可测函数中找出使得期望损失达到最小的预测. 有关损失函数的选取问题的研究可参见文献[9-11]. 本文将以平方损失函数为例,求解未来索赔次数的最优预测. 为了进一步分析先验分布和样本分布对贝叶斯预测产生的影响,根据分布形式是否已知,本文的贝叶斯预测将分为下面 3 种情况进行讨论:(i) 索赔次数的样本分布和风险参数的先验分布均为已知;(ii) 索赔次数的样本分布已知,但风险参数的先验分布是未知的;(iii) 索赔次数的样本分布和风险参数的先验分布均为未知.

本文首先给出保单索赔次数的贝叶斯模型;接着定义索赔次数的贝叶斯预测以及信度预测公式,将根据前面预计讨论的分布形式是否已知的 3 种情况分别求解保单未来可能发生的索赔次数的贝叶斯预测和信度预测,并证明其相关的统计性质;最后,利用数值模拟的方法对索赔次数的这 3 种预测的均方误差进行了比较和分析.

1　索赔次数的贝叶斯模型

考虑 I 个非寿险保单合同,记 M_{ij} 表示第 i 个保单合同在第 j 年的索赔次数,其中 $i = 1,2,\cdots,I$,$j = 1,2,\cdots,n_i$. 这里 n_i 表示第 i 个保单合同的索赔次数

收稿日期:2021-05-20
基金项目:国家自然科学基金(71761019),江西省自然科学基金(20202BABL201001)和江西省教育厅科学技术研究重点课题(GJJ200304)资助项目.
作者简介:章　溢(1985—),女,江西南昌人,讲师,博士,主要从事经济统计与保险精算研究. E-mail:153574268@qq.com

的样本容量. 本文的目的是通过采取适当的方法,根据索赔次数的观测值数据 $\widetilde{M}=\{M_{ij},i=1,2,\cdots,I,j=1,2,\cdots,n_i\}$ 能够更准确地预测未来的索赔次数 $M_{i(n_i+1)}$.

在非寿险保险中,第 i 个保单合同的索赔次数 M_{ij} 的分布常常依赖于某个风险参数 θ_i. 由于风险的非齐次性[1],本文假设 θ_i 为随机变量,它具有某个先验分布,从而建立风险参数的贝叶斯模型.

假设 1 当 θ_i 给定时,索赔次数 $M_{i1},M_{i2},\cdots,M_{in_i}$ 相互独立,且 M_{ij} 的条件分布为

$$P(M_{ij}=x\mid\theta_i)=f_{ij}(x\mid\theta_i),$$

其条件期望和条件方差分别为

$$\mu_{ij}(\theta_i)=E(M_{ij}\mid\theta_i),\sigma_{ij}^2(\theta_i)=\mathrm{Var}(M_{ij}\mid\theta_i).$$

假设 2 风险参数 $\theta_1,\theta_2,\cdots,\theta_I$ 相互独立且具有相同的先验分布 $\pi(\theta)$.

假设 3 各个保单的索赔和风险参数之间相互独立,即 $(\theta_1,\boldsymbol{M}_1^{\mathrm{T}}),(\theta_2,\boldsymbol{M}_2^{\mathrm{T}}),\cdots,(\theta_I,\boldsymbol{M}_I^{\mathrm{T}})$ 相互独立,其中 $\boldsymbol{M}_i=(M_{i1},M_{i2},\cdots,M_{in_i})^{\mathrm{T}}$ 为第 i 个保单的索赔样本.

记 $\mu_{ij}=E(\mu_{ij}(\theta_i))$,$\tau_{ij}^2=\mathrm{Var}(\mu_{ij}(\theta_i))$ 以及 $\sigma_{ij}^2=E(\sigma_{ij}^2(\theta_i))$. μ_{ij}、τ_{ij}^2 和 σ_{ij}^2 都依赖于先验分布 $\pi(\theta)$,它们被称为超参数或结构参数.

下面将根据样本数据 $\widetilde{M}$ 对第 i 个保单在未来 1 年的索赔次数 $M_{i(n_i+1)}$ 进行预测.

记 $T=\{g(\widetilde{M}):g$ 为 $\widetilde{M}$ 的可测函数且 $E_L(g^2(\widetilde{M}))<\infty\}$ 表示样本 $\widetilde{M}$ 的所有可测函数集合. 若 $g(\widetilde{M})$ 为 $M_{i(n_i+1)}$ 的 1 个预测,定义期望预测损失为

$$E_{L_i}(g)=E(L(M_{i(n_i+1)},g(\widetilde{M})))=E((M_{i(n_i+1)}-g(\widetilde{M}))^2),$$

其被称为第 i 个保单的预测风险函数.

定义 1 若有 $g^*\in T$,使得 $E_{L_i}(g^*)\leqslant E_{L_i}(g)$,且至少存在某个 $g_1\in T$,使得 $E_{L_i}(g^*)<E_{L_i}(g_1)$ 成立,则称 g^* 为该损失函数 $L(x,\gamma)$ 下 $M_{i(n_i+1)}$ 的最优预测,简称贝叶斯预测,记为 $\hat{M}_i^B$.

为了得到 $M_{i(n_i+1)}$ 的贝叶斯预测,需要下面的引理,其证明可参见文献[12].

引理 1 在预测 $M_{i(n_i+1)}$ 的过程中,最小化 $E_{L_i}(g)$ 等价于 $\inf\limits_{g\in T}E((M_{i(n_i+1)}-g(\widetilde{M}))^2\mid\widetilde{M})$.

即只需在 $\widetilde{M}$ 给定条件下对 $M_{i(n_i+1)}$ 的预测分布求解最小期望损失,可得到 $M_{i(n_i+1)}$ 的贝叶斯预测.

显然,为求得 $M_{i(n_i+1)}$ 的贝叶斯预测 $\hat{M}_i^B$,需要得到 $M_{i(n_i+1)}$ 的预测分布 $P(M_{i(n_i+1)}=k\mid\widetilde{M})$. 而该预测分布依赖于 M_{ij} 的条件分布 $f_{ij}(x\mid\theta_i)$ 和风险参数 θ_i 的先验分布 $\pi(\theta)$ 的具体分布形式. 为此,本文根据条件分布 $f_{ij}(x\mid\theta_i)$ 和先验分布 $\pi(\theta)$ 是否已知将预测问题分为下面 3 种情况来研究:(i) $f_{ij}(x\mid\theta_i)$ 和先验分布 $\pi(\theta)$ 均为已知;(ii) $f_{ij}(x\mid\theta_i)$ 已知但先验分布 $\pi(\theta)$ 未知;(iii) $f_{ij}(x\mid\theta_i)$ 和先验分布 $\pi(\theta)$ 均为未知.

2 索赔次数的贝叶斯预测和信度预测

本部分将探讨在样本条件分布和先验分布已知或未知时索赔次数的贝叶斯预测问题. 为了模型的简单,样本条件分布取为泊松分布,而风险参数的先验分布取为伽马分布. 泊松分布是刻画离散型随机变量的重要分布,在非寿险精算中有重要的应用[13-16]. 显然,该模型容易推广到其他离散型分布情形,如二项分布和负二项分布模型等.

2.1 在泊松-伽马模型下索赔次数的预测

假设 θ_i 给定,在一定时期内的索赔次数 M_{i1},$M_{i2},\cdots,M_{in_i},\cdots$ 相互独立,且 M_{ij} 服从参数为 $w_{ij}\theta_i$ 的泊松分布,其概率分布律为

$$f_{ij}(x\mid\theta_i)=P(M_{ij}=x\mid\theta_i)=(w_{ij}\theta_i)^x\mathrm{e}^{-w_{ij}\theta_i}/x!,$$
$$x=0,1,2,\cdots,i=1,2,\cdots,I,j=1,2,\cdots,n_i,\quad(1)$$

其中 w_{ij} 为已知的风险暴露数,而参数 $\theta_1,\theta_2,\cdots,\theta_I$ 相互独立且服从伽马分布,其共同的密度函数为

$$\pi(\theta)=\beta^\alpha\theta^{\alpha-1}\mathrm{e}^{-\beta\theta}/(\Gamma(\alpha)),\theta>0.\quad(2)$$

称满足式(1)和式(2)的模型为泊松-伽马模型. 因此得到下面的定理.

定理 1 在泊松-伽马模型中,在给定 $\widetilde{M}=\widetilde{m}$ 的条件下第 i 个保单的未来索赔次数 $M_{i(n_i+1)}$ 的预测分布为负二项分布 $NB(r_i,p_i)$,其中 r_i 和 p_i 分别为

$$r_i=m_{i\cdot}+\alpha,p_i=(\beta+w_{i\cdot})/(\beta+w_{i\cdot}+w_{i(n_i+1)}),$$

这里 $m_{i\cdot}=\sum\limits_{j=1}^{n_i}m_{ij},w_{i\cdot}=\sum\limits_{j=1}^{n_i}w_{ij}$.

证 显然,索赔次数 $M_{i(n_i+1)}$ 是取非负整数值的随机变量,根据连续型全概率公式得到

$$P(M_{i(n_i+1)}=k\mid\widetilde{M}=\widetilde{m})=\int_0^\infty P(M_{i(n_i+1)}=k\mid\theta_i=\theta,\widetilde{M}=\widetilde{m})\pi_{\theta_i}(\theta\mid\widetilde{M}=\widetilde{m})\mathrm{d}\theta.$$

其中 $\pi_{\theta_i}(\theta_i\mid\widetilde{M}=\widetilde{m})$ 为 θ_i 的后验分布,有

$$\pi_{\theta_i}(\theta_i \mid \widetilde{M}=\widetilde{m}) \propto \pi(\theta_i)\prod_{s=1}^{I}\prod_{t=1}^{n_s}(P(M_{st}=m_{st}\mid \theta_s)) \propto \theta_i^{\alpha-1}\mathrm{e}^{-\beta\theta_i}\prod_{s=1}^{I}\prod_{t=1}^{n_s}((w_{st}\theta_s)^{m_{st}}\mathrm{e}^{-w_{st}\theta_s}/m_{st}!) \propto \theta_i^{m_{i\cdot}+\alpha-1}\cdot \mathrm{e}^{-(\beta+w_{i\cdot})\theta_i}.$$

因此,风险参数 θ_i 的后验分布为 Gamma$(m_{i\cdot}+\alpha,\beta+w_{i\cdot})$,其后验密度为

$$\pi_{\theta_i}(\theta \mid \widetilde{M}=\widetilde{m}) = (\beta+w_{i\cdot})^{m_{i\cdot}+\alpha}\theta^{m_{i\cdot}+\alpha-1}\cdot \mathrm{e}^{-(\beta+w_{i\cdot})\theta}/\Gamma(m_{i\cdot}+\alpha),\theta>0.$$

又因为在 θ_i 给定下 $\widetilde{M}$ 与 $M_{i(n_i+1)}$ 相互独立,且有

$$P(M_{i(n_i+1)}=k \mid \theta_i=\theta,\widetilde{M}=\widetilde{m}) = (w_{i(n_i+1)}\theta)^k\cdot \mathrm{e}^{-w_{i(n_i+1)}\theta}/k!,k=0,1,2,\cdots.$$

因此

$$P(M_{i(n_i+1)}=k \mid \widetilde{M}=\widetilde{m}) = \int_0^\infty (w_{i(n_i+1)}\theta)^k\mathrm{e}^{-w_{i(n_i+1)}\theta}/(\beta+w_{i\cdot})^{m_{i\cdot}+\alpha}\theta^{m_{i\cdot}+\alpha-1}\mathrm{e}^{-(\beta+w_{i\cdot})\theta}/(\Gamma(m_{i\cdot}+\alpha)k!)\,\mathrm{d}\theta = (\beta+w_{i\cdot})^{m_{i\cdot}+\alpha}/(\Gamma(m_{i\cdot}+\alpha)k!)\int_0^\infty \mathrm{e}^{-w_{i(n_i+1)}\theta}\theta^{k+m_{i\cdot}+\alpha-1}\cdot \mathrm{e}^{-(\beta+w_{i\cdot})\theta}\mathrm{d}\theta = ((\beta+w_{i\cdot})^{m_{i\cdot}+\alpha}w_{i(n_i+1)}^k/(\Gamma(m_{i\cdot}+\alpha)k!))\cdot(\Gamma(k+m_{i\cdot}+\alpha)/(\beta+w_{i\cdot}+w_{i(n_i+1)})^{k+m_{i\cdot}+\alpha}) = (\Gamma(k+m_{i\cdot}+\alpha)/(\Gamma(m_{i\cdot}+\alpha)k!))((\beta+w_{i\cdot})/(\beta+w_{i\cdot}+w_{i(n_i+1)}))^{m_{i\cdot}+\alpha}\cdot(w_{i(n_i+1)}/(\beta+w_{i\cdot}+w_{i(n_i+1)}))^k.$$

因此,$M_{i(n_i+1)}$ 的预测分布为负二项分布 $NB(r_i,p_i)$.

根据定理1可得到 $M_{i(n_i+1)}$ 的预测分布的矩母函数为

$$G_{M_{i(n_i+1)}\mid\widetilde{M}} = E(\exp(tM_{i(n_i+1)}\mid \widetilde{M}=\widetilde{m})) = ((\beta+w_{i\cdot})/(\beta+w_{i\cdot}+w_{i(n_i+1)}(1-\mathrm{e}^t)))^{m_{i\cdot}+\alpha}.$$

相应地,预测分布的均值和方差分别为

$$E(M_{i(n_i+1)}\mid\widetilde{M}) = (M_{i\cdot}+\alpha)w_{i(n_i+1)}/(\beta+w_{i\cdot}),$$

$$\mathrm{Var}(M_{i(n_i+1)}\mid\widetilde{M}) = (M_{i\cdot}+\alpha)(\beta+w_{i\cdot}+w_{i(n_i+1)})\cdot w_{i(n_i+1)}/(\beta+w_{i\cdot})^2,$$

其中 $m_{i\cdot}=\sum_{j=1}^{n_i}m_{ij}$ 为 $M_{i\cdot}=\sum_{j=1}^{n_i}M_{ij}$ 的观测值.

定理2　在平方损失函数中,第 i 个保单的索赔次数 $M_{i(n_i+1)}$ 的贝叶斯估计恰好为预测均值,即有

$$\hat{M}_i^B = (M_{i\cdot}+\alpha)w_{i(n_i+1)}/(\beta+w_{i\cdot}), \quad (3)$$

且最小均方误差为

$$E((\hat{M}_i^B - M_{i(n_i+1)})^2) = (w_{i\cdot}\theta_0+\alpha)(\beta+w_{i\cdot}+w_{i(n_i+1)})w_{i(n_i+1)}/(\beta+w_{i\cdot})^2.$$

证　令 $\Phi = E((M_{i(n_i+1)}-g(\widetilde{M}))^2\mid\widetilde{M})$. 则在 $\widetilde{M}$ 给定条件下,$g(\widetilde{M})$ 是一个给定的常数. 由定理1知,只需求 $g(\widetilde{M})$ 使得 Φ 达到最小. 因此,对 Φ 关于 $g(\widetilde{M})$ 求导,并令导数为零,得到 $g(\widetilde{M}) = E(M_{i(n_i+1)}\mid\widetilde{M})$. 且有 $\partial^2\Phi/\partial g^2 = -2<0$. 因此,$M_{i(n_i+1)}$ 的贝叶斯预测为 $\hat{M}_i^B = E(M_{i(n_i+1)}\mid\widetilde{M})$. 由定理1得到式(3). 注意到 $E(M_{i\cdot}) = w_{i\cdot}\theta_0$,由双重期望公式知,贝叶斯预测 $\hat{M}_i^{B_1}$ 的均方损失为

$$E((\hat{M}_i^B - g(\widetilde{M}))^2) = E\{E((E(M_{i(n_i+1)}\mid\widetilde{M})-g(\widetilde{M}))^2\mid\widetilde{M}\} = E(\mathrm{Var}(M_{i(n_i+1)}\mid\widetilde{M})) = E((M_{i\cdot}+\alpha)(\beta+w_{i\cdot}+w_{i(n_i+1)})w_{i(n_i+1)}/(\beta+w_{i\cdot})^2) = \alpha(\beta+w_{i\cdot}+w_{i(n_i+1)})w_{i(n_i+1)}/(\beta(\beta+w_{i\cdot})).$$

注1　注意到贝叶斯预测 $\hat{M}_i^{B_1}$ 可以表达为下面的加权和的形式:

$\hat{M}_i^B = Z_{Bi}(n_i w_{i(n_i+1)}\overline{M}_i/w_{i\cdot}) + (1-Z_{Bi})\alpha w_{i(n_i+1)}/\beta$,其中 $Z_{Bi} = w_{i\cdot}/(\beta+w_{i\cdot})$ 为权重因子. 注意到 $E(M_{i(n_i+1)}) = \alpha w_{i(n_i+1)}/\beta$,因此贝叶斯预测 $\hat{M}_i^{B_1}$ 表达为聚合均值 $E(M_{i(n_i+1)})$ 和样本的修正均值 $n_i w_{i(n_i+1)}\cdot\overline{M}_i/w_{i\cdot}$ 的加权平均. 由于 $w_{ij}>0$,所以权重 Z_{Bi} 是样本容量 n_i 的增函数.

注2　在定理2中,若 $w_{ij}=1$,则索赔次数 $M_{i(n_i+1)}$ 的贝叶斯预测退化为

$$\hat{M}_{ii}^B = (M_{i\cdot}+\alpha)/(\beta+n_i),$$

其预测均方误差为

$$E((\hat{M}_{ii}^B - M_{i(n_i+1)})^2) = \alpha(\beta+n_i+1)/(\beta(\beta+n_i)).$$

此时 n_i 为样本容量,当 $n_i\to\infty$ 时,有

$$\hat{M}_{ii}^B = (M_{i\cdot}+\alpha)/(\beta+n_i) = (\overline{M}_i+\alpha/n_i)/(\beta/n_i+1)\to\theta_i,\text{a.s.}, \quad (4)$$

$$\lim_{n_i\to\infty}E((\hat{M}_{ii}^B - M_{i(n_i+1)})^2) = \lim_{n_i\to\infty}(\alpha(\beta+n_i+1)/(\beta(\beta+n_i))) = \alpha/\beta.$$

这说明 $\hat{M}_{ii}^B$ 的均方预测误差收敛到一个常数 α/β. 由于 $M_{i(n_i+1)}$ 是随机变量,所以其均方预测误差可分解为

$$E((\hat{M}_{ii}^B - M_{i(n_i+1)})^2) = E((\hat{M}_{ii}^B-\theta_i)^2) + E((\theta_i - M_{i(n_i+1)})^2). \quad (5)$$

根据式(4)知 $\hat{M}_{ii}^B$ 是 θ_i 的强相合估计,则均方预测误差式(5)的第1部分为

$$E((\hat{M}_{ii}^B-\theta_i)^2) = \{E((\hat{M}_{ii}^B-\theta_i)^2\mid\widetilde{M})\} = E(\mathrm{Var}(\theta_i\mid\widetilde{M})) = \alpha/(\beta(\beta+n_i))\to 0,$$

而均方预测误差式(5)的第2部分

$$E\left(\left(\theta_i - M_{i(n_i+1)}\right)^2\right) = E\left\{E\left(\left(\theta_i - M_{i(n_i+1)}\right)^2 \mid \theta_i\right)\right\} = E\left(\mathrm{Var}\left(M_{i(n_i+1)} \mid \theta_i\right)\right) = \alpha/\beta.$$

因此,均方预测误差收敛到常数 α/β 主要是由 $M_{i(n_i+1)}$ 的随机性导致的.

在实际运用中,超参数 α 与 β 一般是未知的,需要根据样本来估计. 常用的估计方法包括 2 种: 一种是极大边际极大似然估计,另一种是矩法估计. 本节仅考虑极大似然估计的方法估计超参数. 注意到样本 $\widetilde{M} = \{M_{ij}, i = 1,2,\cdots,I, j = 1,2,\cdots,n_i\}$ 的联合分布为

$$\int_0^{\infty}\prod_{i=1}^{I}\left(\left(\prod_{j=1}^{n_i}P\left(M_{ij} = m_{ij} \mid \theta_i\right)\right)\pi\left(\theta_i\right)\mathrm{d}\theta_1\cdots\mathrm{d}\theta_I = \int_0^{\infty}\prod_{i=1}^{I}\left(\left(\prod_{j=1}^{n_i}\left(\left(w_{ij}\theta_i\right)^{m_{ij}}\beta^{\alpha}\mathrm{e}^{-w_{ij}\theta_i}/\left(m_{ij}!\,\Gamma\left(\alpha\right)\right)\right)\theta_i^{\alpha-1}\cdot\mathrm{e}^{-\beta\theta_i}\right)\mathrm{d}\theta_1\cdots\mathrm{d}\theta_I = \int_0^{\infty}\cdots\int_0^{\infty}\prod_{i=1}^{I}\left(\left(\prod_{j=1}^{n_i}w_{ij}\right)\beta^{\alpha}\theta_i^{\alpha+m_{i\cdot}-1}\cdot\mathrm{e}^{-(\beta+w_{i\cdot})\theta_i}/\left(m_{i1}!\cdots m_{in_i}!\,\Gamma\left(\alpha\right)\right)\right)\mathrm{d}\theta_1\cdots\mathrm{d}\theta_I = \left(\prod_{i=1}^{I}\prod_{j=1}^{n_i}w_{ij}\right)\cdot\beta^{I\alpha}\prod_{i=1}^{I}\Gamma\left(\alpha + m_{i\cdot}\right)/\left(\left(\prod_{i=1}^{I}\prod_{j=1}^{n_i}m_{ij}!\right)\left(\Gamma\left(\alpha\right)\right)^{I}\prod_{i=1}^{I}\left(\beta + w_{i\cdot}\right)^{\alpha+m_{i\cdot}}\right).$$

因此,超参数 α 与 β 的对数似然函数为

$$l\left(\alpha,\beta\right) = C + \sum_{i=1}^{I}\left(\ln\Gamma\left(\alpha + m_{i\cdot}\right) - \ln\Gamma\left(\alpha\right)\right) + I\alpha\ln\beta - \sum_{i=1}^{I}\left(\alpha + m_{i\cdot}\right)\ln\left(\beta + w_{i\cdot}\right).$$

令

$$\begin{cases}\partial l\left(\alpha,\beta\right)/\partial\alpha = 0,\\ \partial l\left(\alpha,\beta\right)/\partial\beta = 0,\end{cases} \tag{6}$$

则可得到 α 与 β 的极大似然估计. 显然,极大似然估计一般没有显式表达式. 需要根据迭代算法得到极大似然估计的近似解.

记 $\hat{\alpha}$ 与 $\hat{\beta}$ 分别为 α 与 β 的估计,代入后得到未来索赔次数的经验贝叶斯预测为

$$\widetilde{M}_i^B = \left(M_{i\cdot} + \hat{\alpha}\right)w_{i(n_i+1)}/\left(\hat{\beta} + w_{i\cdot}\right). \tag{7}$$

此时,经验贝叶斯预测 $\widetilde{M}_i^B$ 不依赖于任何未知参数,可以在实际中直接用于保险索赔次数的预测.

2.2 当先验分布未知时索赔次数的信度预测

在 2.1 节中,研究了在泊松-伽马模型中未来索赔次数的贝叶斯预测. 显然,贝叶斯预测的获取依赖于索赔次数的样本分布和风险参数的先验分布具体分布形式. 但是,在保险的实际应用中,风险参数的先验分布往往是未知的. 此时,索赔次数的贝叶斯预测没有显示表达式. 在非寿险精算中,H. Bühlmann 等[17] 将风险保费的估计限定在样本的线性函数中,得到了风险保费的信度估计. 信度理论不仅在非寿险精算的保费定价中有重要的应用[18-20],而且被用于责任准备金评估、时间序列预测等方面[21-23].

假设在 θ_i 给定条件下,索赔次数 $M_{i1}, M_{i2}, \cdots, M_{in_i}, \cdots$ 相互独立,且 M_{ij} 服从参数为 $w_{ij}\theta_i$ 的泊松分布,其概率分布律为式(1). 而参数 $\theta_1, \theta_2, \cdots, \theta_I$ 相互独立且服从某个未知的先验分布 $\pi(\theta)$. 称该模型为泊松-未知先验分布模型. 本节的目的仍然是在给定损失函数中求解 $M_{i(n_i+1)}$ 的最优预测.

在平方损失函数中,将 M_{i,n_i+1} 的预测限定在样本的线性函数中,求解期望平方损失达到最小的最优预测. 定义下面的线性函数预测集合:

$$L\left(\widetilde{M},1\right) = \left\{a_0 + \sum_{s=1}^{I}\sum_{t=1}^{n_s}a_{st}M_{st}\,\middle|\,a_0, a_{st} \in \mathbf{R}\right\}.$$

并求解最小化问题

$$\min_{g\in L\left(\widetilde{M},1\right)}E\left(\left(M_{i,n_i+1} - g\right)^2\right) = \min_{a_0,a_{st}\in\mathbf{R}}E\left(\left(M_{i,n_i+1} - a_0 - \sum_{s=1}^{I}\sum_{t=1}^{n_s}a_{st}M_{st}\right)^2\right). \tag{8}$$

根据泊松分布的性质,有

$$E\left(M_{ij} \mid \theta_i\right) = \mathrm{Var}\left(M_{ij} \mid \theta_i\right) = w_{ij}\theta_i.$$

令 $Y_{ij} = M_{ij}/m_{ij}$,则 $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}, \cdots$ 相互独立,且

$$E\left(Y_{ij} \mid \theta_i\right) = \theta_i, \mathrm{Var}\left(Y_{ij} \mid \theta_i\right) = \theta_i/w_{ij}.$$

记 $E\left(\theta_i\right) = \theta_0, \mathrm{Var}\left(\theta_i\right) = \gamma$,则有 $\mathrm{Var}\left(E\left(Y_{ij} \mid \theta_i\right)\right) = \gamma, E\left(w_{ij}\mathrm{Var}\left(Y_{ij} \mid \theta_i\right)\right) = \theta_0, E\left(Y_{ij}\right) = \theta_0$.

根据信度理论的求解,可得到下面的结论.

定理 3 在泊松-未知先验分布模型中,求解最小化期望平方损失式(8)得到 $M_{i(n_i+1)}$ 的最优线性预测为 $\hat{M}_i^C = Z_{Ci}\left(n_i w_{i(n_i+1)}\overline{M}_i/w_{i\cdot}\right) + \left(1 - Z_{Ci}\right)w_{i(n_i+1)}\theta_0$,其中 $\overline{M}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i}M_{ij}$ 为第 i 个保单的索赔次数样本均值,且 $Z_{Ci} = w_{i\cdot}\gamma/\left(w_{i\cdot}\gamma + \theta_0\right)$.

证 根据文献[7],最小化问题(8)的解为 $M_{i(n_i+1)}$ 在样本生成的线性空间上的投影,则有

$$\hat{M}_i^{Ci} = E\left(M_{i(n_i+1)}\right) + \mathrm{Cov}\left(M_{i(n_i+1)}, \boldsymbol{M}\right)\left(\mathrm{Var}\left(\boldsymbol{M}\right)\right)^{-1}\cdot\left(\boldsymbol{M} - E\left(\boldsymbol{M}\right)\right),$$

其中 $\boldsymbol{M} = \left(M_{11}, M_{12}, \cdots, M_{1,n_1}, \cdots, M_{I1}, M_{I2}, \cdots, M_{In_i}\right)^{\mathrm{T}}$. 暂时记

$$\boldsymbol{u}^{\mathrm{T}} = \left(M_{i1}, M_{i2}, \cdots, M_{in_i}\right), \boldsymbol{w}^{\mathrm{T}} = \left(w_{i1}, w_{i2}, \cdots, w_{in_i}\right), \boldsymbol{V} = \begin{pmatrix}1/w_{i1} & & \\ & \ddots & \\ & & 1/w_{in_i}\end{pmatrix}.$$

根据风险之间的独立性,有

$$\hat{M}_i^C = E(M_{i(n_i+1)}) + \mathrm{Cov}(M_{i(n_i+1)}, \boldsymbol{u})(\mathrm{Var}(\boldsymbol{u}))^{-1} \cdot (\boldsymbol{u} - E(\boldsymbol{u})).$$

注意到$E(M_{i(n_i+1)}) = w_{i(n_i+1)}\theta_0$,$E(\boldsymbol{u}) = \boldsymbol{w}\theta_0$,$\mathrm{Cov}(M_{i(n_i+1)}, \boldsymbol{u}) = w_{i(n_i+1)}\boldsymbol{w}^{\mathrm{T}}\gamma$ 以及

$$(\mathrm{Var}(\boldsymbol{u}))^{-1} = (\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\gamma + \boldsymbol{V}^{-1}\theta_0)^{-1} = \boldsymbol{V}/\theta_0 - (\boldsymbol{V}/\theta_0)\boldsymbol{w}(1/\gamma + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{w}/\theta_0)^{-1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}/\theta_0 = (\boldsymbol{V} - \gamma\boldsymbol{V}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}/(w_{i\cdot}\gamma + \theta_0))/\theta_0.$$

因此有

$$\hat{M}_i^C = w_{i(n_i+1)}\theta_0 + w_{i(n_i+1)}\boldsymbol{w}^{\mathrm{T}}\gamma(\boldsymbol{V} - \gamma\boldsymbol{V}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}/(w_{i\cdot}\gamma + \theta_0))(u - w\theta_0)/\theta_0 = w_{i(n_i+1)}(\theta_0 + \gamma\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}(u - w\theta_0)/(w_{i\cdot}\gamma + \theta_0)) = w_{i(n_i+1)}(\theta_0 + w_{i\cdot}\gamma/(w_{i\cdot}\gamma + \theta_0) \cdot (\sum_{j=1}^{n_i} M_{ij}/w_{i\cdot} - \theta_0)) = w_{i(n_i+1)}(n_i\gamma\overline{M}_i/(w_{i\cdot}\gamma + \theta_0) + \theta_0^2/(w_{i\cdot}\gamma + \theta_0)) = Z_{Ci}n_iw_{i(n_i+1)}\overline{M}_i/w_{i\cdot} + (1 - Z_{Ci}) \cdot w_{i(n_i+1)}\theta_0.$$

因此称 $\hat{M}_i^C$ 为 $M_{i(n_i+1)}$ 的信度预测,其中称 Z_{Ci} 为第 i 个保单的信度因子.

在泊松-未知先验分布模型中,由于先验分布 $\pi(\theta)$ 是未知的,因此无法求得样本 $\widetilde{M} = \{M_{ij}, i = 1, 2, \cdots, I, j = 1, 2, \cdots, n_i\}$ 的联合分布,因此不能得到未知参数 γ 和 θ_0 的似然函数,从而极大似然估计无法直接使用. 这时可考虑 γ 和 θ_0 的矩估计.

令

$$\overline{Y}_i^w = \frac{1}{w_{i\cdot}}\sum_{j=1}^{n_i} w_{ij}Y_{ij} = n_i\overline{M}_i/w_{i\cdot},$$

$$\overline{\overline{Y}}^w = \frac{1}{w_{\cdot\cdot}}\sum_{j=1}^{n_i} w_{i\cdot}\overline{Y}_i^w = \frac{1}{w_{\cdot\cdot}}\sum_{i=1}^{I}\sum_{j=1}^{n_i} M_{ij},$$

$$T = \sum_{i=1}^{I} w_{i\cdot}(\overline{Y}_i^w - \overline{\overline{Y}}^w)^2,$$

则 $E(\overline{Y}_i^w \mid \theta_i) = \theta_i$,$\mathrm{Var}(\overline{Y}_i^w \mid \theta_i) = \theta_i/w_{i\cdot}$,因此 $E(\overline{\overline{Y}}^w) = \theta_0$. 则 θ_0 的一个无偏估计为

$$\hat{\theta}_0 = \overline{\overline{Y}}^w = \frac{1}{w_{\cdot\cdot}}\sum_{i=1}^{I}\sum_{j=1}^{n_i} M_{ij}.$$

注意到 $E(\overline{Y}_i^w - \overline{\overline{Y}}^w) = 0$,$\mathrm{Var}(\overline{Y}_i^w) = \theta_0/w_{i\cdot} + \gamma$ 以及

$$\mathrm{Var}(\overline{\overline{Y}}^w) = \theta_0/w_{\cdot\cdot} + \frac{\gamma}{w_{\cdot\cdot}^2}\sum_{k=1}^{I} w_k^2\cdot\mathrm{Cov}(\overline{Y}_i^w, \overline{\overline{Y}}^w) = \theta_0/w_{\cdot\cdot} + \gamma w_{i\cdot}/w_{\cdot\cdot},$$

则有

$$E(T) = \sum_{i=1}^{I} w_{i\cdot}E(\overline{Y}_i^w - \overline{\overline{Y}}^w)^2 = \sum_{i=1}^{I} w_{i\cdot}\mathrm{Var}(\overline{Y}_i^w - \overline{\overline{Y}}^w) = \sum_{i=1}^{I} w_{i\cdot}(\mathrm{Var}(\overline{Y}_i^w) + \mathrm{Var}(\overline{\overline{Y}}^w) - 2\mathrm{Cov}(\overline{Y}_i^w, \overline{\overline{Y}}^w)) = \sum_{i=1}^{I} w_{i\cdot}(\theta_0/w_{i\cdot} + \gamma + \theta_0/w_{\cdot\cdot} + \frac{\gamma}{w_{\cdot\cdot}^2}\sum_{k=1}^{I} w_{k\cdot}^2 - 2(\theta_0/w_{\cdot\cdot} + \gamma w_{i\cdot}/w_{\cdot\cdot})) = (I - 1)\theta_0 + \frac{\gamma}{w_{\cdot\cdot}}(w_{\cdot\cdot}^2 - \sum_{k=1}^{I} w_{k\cdot}^2).$$

因此可以得到 γ 的一个无偏估计为

$$\hat{\gamma} = \frac{w_{\cdot\cdot}}{w_{\cdot\cdot}^2 - \sum_{k=1}^{I} w_{k\cdot}^2}(\sum_{i=1}^{I} w_{i\cdot}(\overline{Y}_i^w - \overline{\overline{Y}}^w)^2 - (I - 1)\overline{\overline{Y}}^w).$$

在实际运用中,一般取 $\hat{\gamma}^* = \max(\hat{\gamma}, 0)$ 作为 γ 的估计. 将结构参数的估计 $\hat{\theta}_0$ 和 $\hat{\gamma}^*$ 代入后获得未来索赔次数的经验贝叶斯预测

$$\hat{M}_i^C = \hat{Z}_{Ci}(n_iw_{i(n_i+1)}\overline{M}_i/w_{i\cdot}) + (1 - \hat{Z}_{Ci})w_{i(n_i+1)}\hat{\theta}_0, \quad (9)$$

其中

$$\hat{Z}_{Ci} = n_i\hat{\gamma}^*/(n_i\hat{\gamma}^* + \hat{\theta}_0). \quad (10)$$

类似式(10),称 $\widetilde{M}_i^{Ci}$ 为在先验分布未知时未来索赔次数 $M_{i(n_i+1)}$ 的经验贝叶斯信度预测. 在实际运用中,若风险参数 θ_i 的先验分布 $\pi(\theta)$ 是未知的,则经验贝叶斯预测式(7)是无法使用的,而此时可以使用经验贝叶斯信度预测式(9).

2.3　当完全未知分布时索赔次数的预测

在实践中,有可能索赔次数 M_{ij} 的条件分布和 θ_i 的先验分布都是未知的,这时需要预测未来索赔额.

假设在 θ_i 给定条件下,索赔次数 $M_{i1}, M_{i2}, \cdots, M_{in_i}, \cdots$ 相互独立且服从某个概率分布律 $P(M_{ij} = x \mid \theta_i) = f_{ij}(x; \theta_i)$. 而 $\theta_1, \theta_2, \cdots, \theta_I$ 相互独立且服从共同的先验分布 $\pi(\theta)$. 为了与未知先验分布的泊松模型进行对比,进一步假设索赔次数 M_{ij} 的条件期望和条件方差分别为

$$E(M_{ij} \mid \theta_i) = w_{ij}\theta_i, \mathrm{Var}(M_{ij} \mid \theta_i) = w_{ij}\sigma^2(\theta_i),$$

且记

$$\theta_0 = E(\mu(\theta_i)), \gamma = \mathrm{Var}(\theta_i), \varphi = E(\sigma^2(\theta_i)),$$

称满足上述假设的模型为在未知分布时索赔次数的贝叶斯模型.

定理 4　在未知分布的贝叶斯模型中,求解最小化期望平方损失(8)得到 $M_{i(n_i+1)}$ 的最优线性预测为 $\hat{M}_i^D = Z_{Di}n_iw_{i(n_i+1)}\overline{M}_i/w_{i\cdot} + (1 - Z_{Di})w_{i(n_i+1)}\theta_0$,其中 $\overline{M}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} M_{ij}$ 为第 i 个保单的索赔次数样本均

值,而权重 Z_{Di} 为 $Z_{Di}=w_{i\cdot}\gamma/(w_{i\cdot}\gamma+\varphi)$.

证 本定理的证明类似于定理 3.

当结构参数 θ_0、γ 和 φ 未知时,可构造它们的估计,其构造方法类似于文献[7].得到

$$\hat{\theta}_0=\bar{\bar{Y}}^{w}=\frac{1}{w_{\cdot\cdot}}\sum_{i=1}^{I}\sum_{j=1}^{n_i}M_{ij},$$

$$\hat{\varphi}=\frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i^{w}\right)^2\right),$$

$$\hat{\gamma}_1=\frac{w_{\cdot\cdot}}{w_{\cdot\cdot}^2-\sum_{k=1}^{I}w_{k\cdot}^2}\left(\sum_{i=1}^{I}w_{i\cdot}\left(\bar{Y}_i^{w}-\bar{\bar{Y}}^{w}\right)^2-(I-1)\hat{\varphi}\right).$$

容易证明,在未知分布的索赔次数贝叶斯模型中,$\hat{\theta}_0$、$\hat{\gamma}_1$ 和 $\hat{\varphi}$ 均是各自参数的无偏估计.将结构参数的估计代入,得到未来索赔次数的经验贝叶斯估计为

$$\widetilde{M}_i^{D}=\hat{Z}_{Di}n_iw_{i(n_i+1)}\bar{M}_i/w_{i\cdot}+\left(1-\hat{Z}_{Di}\right)w_{i(n_i+1)}\hat{\theta}_0,\quad(11)$$

其中 $\hat{Z}_{Di}=n_i\hat{\gamma}_1/(n_i\hat{\gamma}_1+\hat{\varphi})$.

类似式(7),称 $\widetilde{M}_i^{D}$ 为在先验分布未知时未来索赔次数 M_{i,n_i+1} 的经验贝叶斯信度预测.在实际中,若风险参数 θ_i 的先验分布 $\pi(\theta)$ 是未知的,则经验贝叶斯预测式(7)是无法使用的,而此时可以使用经验贝叶斯信度预测式(11).

3 数值模拟与比较分析

第 2 节根据索赔次数和风险参数的分布是否已知,给出了 3 种情形的索赔次数预测模型.在保险的实际运用中,希望尽可能利用更多的信息预测未来的索赔次数.因此,当有很强的证据能明确索赔次数或风险参数的具体分布时,2.1 节的模型显然要优于 2.2 节和 2.3 节的模型.然而,若没有证据表明索赔次数或风险参数的具体分布,则 2.3 节的模型是首选.实际情况表明:常常能拟合出样本的具体分布,而先验分布往往是未知的,此时 2.2 节的结果就更具有说服力.下面将用数值模拟的方法来验证这些结论.

在数值模拟中,为了方便,假设 $w_{ij}=1$,且 $n_1=n_2=\cdots=n_I=n$,则在第 1 个模型中未来索赔次数的贝叶斯预测退化为

$$\hat{M}_i^{B}=\left(\sum_{j=1}^{n}M_{ij}+\alpha\right)/(\beta+n).$$

在第 2 个模型中,未来索赔次数的线性贝叶斯预测退化为

$$\hat{M}_i^{C}=Z_{Ci}\bar{M}_i+(1-Z_{Ci})\theta_0,$$

其中 $\theta_0=E(\theta)$,$\bar{M}_i=\sum_{j=1}^{n}M_{ij}/n$,$Z_{Ci}=n\gamma/(n\gamma+\theta_0)$,$\gamma=\mathrm{Var}(\theta)$.

在第 3 个模型中,未来索赔次数的线性贝叶斯预测退化为

$$\hat{M}_i^{D}=Z_{Di}\bar{M}_i+(1-Z_{Di})\theta_0,$$

其中 $Z_D=n\gamma/(n\gamma+\varphi)$.

下面分别讨论在 3 个模型中结构参数的估计.在模型 1 中,极大似然方程(6)退化为

$$\partial l(\alpha,\beta)/\partial\alpha=\sum_{i=1}^{I}\left(\dot{\Gamma}(\alpha+m_{i\cdot})/\Gamma(\alpha+m_{i\cdot})-\dot{\Gamma}(\alpha)/\Gamma(\alpha)-\ln(\beta+n)+\ln\beta\right)=0,$$

$$\partial l(\alpha,\beta)/\partial\beta=I\alpha/\beta-I(\alpha+\bar{\bar{M}})/(\beta+n)=0,$$

即有 $\alpha=\beta\bar{\bar{M}}$,其中 α 由方程

$$\frac{1}{I}\sum_{i=1}^{I}\left(G(\alpha+m_{i\cdot})-G(\alpha)\right)=1+n\bar{\bar{M}}/\alpha$$

所确定,这里 $G(\alpha)=\dot{\Gamma}(\alpha)/\Gamma(\alpha)$,$\bar{\bar{M}}=\sum_{i=1}^{I}\sum_{j=1}^{n}M_{ij}/nI$.

在第 2 和第 3 个模型中,由于先验分布是未知的,因此采用矩估计法估计结构参数.如第 2 个模型的结构参数估计退化为

$$\hat{\theta}_0=\bar{\bar{M}},\hat{\gamma}=\frac{1}{I-1}\sum_{i=1}^{I}\left(\bar{M}_i-\bar{\bar{M}}\right)^2-\bar{\bar{M}}/n,$$

第 3 个模型的结构参数估计退化为

$$\hat{\theta}_0=\bar{\bar{M}},\hat{\varphi}=\frac{1}{I(n-1)}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(M_{ij}-\bar{M}_i\right)^2,$$

$$\hat{\gamma}_1=\frac{1}{I-1}\sum_{i=1}^{I}\left(\bar{M}_i-\bar{\bar{M}}\right)^2-\hat{\varphi}/n.$$

数值模拟分为 3 个部分.首先,假设索赔次数服从泊松分布,风险参数服从伽马分布,分别比较 3 个预测均方误差;其次,假设索赔次数服从泊松分布,而先验分布未知,对 3 个预测的均方预测误差进行比较;最后,假设样本分布和先验分布都是未知的,进而比较 3 个估计的均方预测误差.

注意到 $\hat{M}_i^{k}$ 的均方预测误差为

$$M_{PSE_k}=E\left(\left(\hat{M}_i^{k}-M_{i(n_i+1)}\right)^2\right)=E\left(\left(\hat{M}_i^{k}-\theta_i+\theta_i-M_{i(n_i+1)}\right)^2\right)=E\left(\left(\hat{M}_i^{k}-\theta_i\right)^2\right)+E\left(\left(\theta_i-M_{i(n_i+1)}\right)^2\right)+2E\left(\left(\hat{M}_i^{k}-\theta_i\right)\left(\theta_i-M_{i(n_i+1)}\right)\right)=E\left(\left(\hat{M}_i^{k}-\theta_i\right)^2\right)+E\left(\mathrm{Var}\left(M_{i(n_i+1)}\mid\theta_i\right)\right).$$

在上述均方预测误差中第 2 项不依赖于预测的

表达式,因此只需要比较3 个估计的第1 项. 记 $\hat{M}_i^k$ 的均方预测误差的第1 项为 $M_{SE_k} = E((\hat{M}_i^k - \theta_i)^2)$,其中 $k = B、C、D$. 因此,只需要比较 3 个估计的均方预测误差的第 1 项 M_{SE_k},称之为均方误差.

3.1　泊松-伽马模型

假设在 θ_i 给定条件下 $M_{i1}, M_{i1}, \cdots$ 相互独立且服从参数为 θ_i 的泊松分布,且 θ_i 服从伽马分布 Gamma(α,β),取 $\alpha = 1, \beta = 2, I = 8$. 运用 R 软件模拟 10 000 次,得到表 1 ~ 表 3.

表 1　在泊松-伽马模型中贝叶斯预测的均方误差($n = 10$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$100M_{SE_B}$	4.318 3	4.510 0	3.861 0	4.377 4	4.060 7	4.461 8	4.468 8	4.445 1
$100M_{SE_C}$	4.581 6	4.693 0	4.181 1	4.605 3	4.378 0	4.660 8	4.763 1	4.711 7
$100M_{SE_D}$	4.598 0	4.714 3	4.195 2	4.621 9	4.393 8	4.678 3	4.776 8	4.726 3

表 2　在泊松-伽马模型中贝叶斯预测的均方误差($n = 50$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$100M_{SE_B}$	0.477 0	0.480 9	0.466 2	0.484 3	0.509 8	0.518 0	0.539 1	0.471 5
$100M_{SE_C}$	0.479 8	0.484 6	0.471 4	0.486 2	0.515 0	0.521 2	0.541 9	0.476 8
$100M_{SE_D}$	0.479 8	0.484 6	0.471 4	0.486 2	0.515 0	0.521 2	0.541 9	0.476 8

表 3　在泊松-伽马模型中贝叶斯预测的均方误差($n = 200$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$100M_{SE_B}$	0.242 4	0.257 1	0.247 1	0.245 8	0.232 5	0.248 6	0.261 1	0.244 0
$100M_{SE_C}$	0.243 4	0.257 9	0.248 4	0.246 4	0.233 1	0.249 1	0.261 9	0.245 4
$100M_{SE_D}$	0.243 4	0.257 9	0.248 4	0.246 4	0.233 1	0.249 1	0.261 9	0.245 4

根据表 1 ~ 表 3 的模拟结果可得到以下结论: (i) 在泊松 -伽马模型中,3 个估计的均方误差的大小关系为 $M_{SE_B} < M_{SE_C} < M_{SE_D}$. 显然,由于预测 $\hat{M}_i^B$ 与第 1 个模型的假设相符,因此有最小的均方误差. 然而 $\hat{M}_i^C$ 仅仅使用了泊松分布的分布信息,却忽略了伽马先验分布的信息,导致均方误差有所增加; $\hat{M}_i^D$ 没有利用任何分布信息,因此均方误差最大,但随着样本容量的增大,3 个预测的均方预测误差的差异越来越小(当 $n \geqslant 100$ 时, $\hat{M}_i^C$ 和 $\hat{M}_i^D$ 的均方误差几乎相等). (ii) 3 个估计的均方误差 M_{SE_k} 都能有收敛到 0 的趋势,这说明 3 个预测都是风险参数的相合估计.

3.2　泊松-均匀分布模型

假设在 θ_i 给定条件下 $M_{i1}, M_{i2}, \cdots$ 相互独立且服从参数为 θ_i 的泊松分布,且 θ_i 服从对数正态 $U(0,1)$ 分布. 为了与模型 1 进行比较,对 $I = 8$ 及不同的样本容量,计算 3 个预测的均方误差

$$M_{SE_k} = E((\hat{M}_i^k - \theta_i)^2),$$

其中 $k = B、C、D$,得到表 4 ~ 表 6.

表 4　在泊松-均匀模型中贝叶斯预测的均方误差($n = 10$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$100M_{SE_B}$	3.982 5	3.901 0	3.972 3	3.864 1	3.904 5	3.822 1	3.773 8	3.932 3
$100M_{SE_C}$	3.803 2	3.923 0	3.845 2	3.811 5	3.845 6	3.692 1	3.722 0	3.881 7
$100M_{SE_D}$	3.818 8	3.935 5	3.860 5	3.825 8	3.860 1	3.706 1	3.735 2	3.896 2

表 5　在泊松-均匀模型中贝叶斯预测的均方误差($n = 100$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$100M_{SE_B}$	0.476 3	0.494 9	0.484 0	0.462 7	0.475 8	0.495 1	0.484 9	0.478 5
$100M_{SE_C}$	0.472 0	0.491 2	0.480 0	0.462 1	0.474 1	0.491 0	0.483 8	0.475 7
$100M_{SE_D}$	0.472 1	0.491 3	0.480 1	0.462 2	0.474 1	0.491 0	0.483 9	0.475 8

表 6 在泊松-均匀模型中贝叶斯预测的均方误差($n = 200$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$100M_{SE_B}$	0.241 8	0.249 1	0.247 8	0.244 3	0.242 7	0.250 2	0.247 9	0.244 9
$100M_{SE_C}$	0.241 7	0.247 0	0.248 5	0.242 5	0.242 7	0.248 7	0.246 2	0.243 4
$100M_{SE_D}$	0.241 7	0.247 0	0.248 5	0.242 5	0.242 8	0.248 7	0.246 2	0.243 4

在泊松-均匀分布模型中,贝叶斯预测 $\hat{M}_i^B$ 由于对先验分布的假设错误导致均方误差增大,这时均方误差的大小关系为 $M_{SE_C} < M_{SE_B} < M_{SE_D}$. 预测 $\hat{M}_i^C$ 不仅利用了泊松分布的信息,而且对先验分布不假设具体分布,因而均方误差最小. $\hat{M}_i^D$ 虽然没有模型假设的错误,但也忽略了泊松分布的信息,导致估计误差比 $\hat{M}_i^C$ 的均方误差稍许大一些. 当样本容量为 200 时,预测 $\hat{M}_i^C$ 和 $\hat{M}_i^D$ 的均方误差几乎相等.

3.3 二项-贝塔分布模型

假设在 θ_i 给定条件下 $X_{i1}, X_{i2}, \cdots$ 相互独立且服从参数为 θ_i 的二项分布 $B(m, \theta_i)$,其概率分布为

$$P(X_{ij} = k \mid \theta_i) = \binom{m}{k}(1-\theta_i)^{m-k}\theta_i^k, k = 0,1,2,\cdots,m.$$

且假设 θ_i 服从 $\mathrm{Beta}(\alpha,\beta)$ 分布. 为了与前面 2 个模型比较,令 $M_{ij} = X_{ij}/m$,且取 $m = 50$,$a = b = 3$,对 $I = 8$ 和不同样本容量,计算 3 个预测的均方预测误差

$$M_{SE_k} = E((\hat{M}_i^k - \theta_i)^2),$$

其中 $k = B、C、D$,得到表 7 ~ 表 9.

表 7 在二项-贝塔分布模型中贝叶斯预测的均方误差($n = 10$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$1\,000M_{SE_B}$	32.027 6	24.530 7	26.657 2	27.374 0	28.265 5	27.043 8	27.319 4	27.760 2
$1\,000M_{SE_C}$	11.416 4	9.045 3	9.756 6	10.076 8	10.096 8	9.754 3	10.004 7	10.143 1
$1\,000M_{SE_D}$	0.415 7	0.432 4	0.431 7	0.430 2	0.424 5	0.425 7	0.428 5	0.424 8

表 8 在二项-贝塔分布模型中贝叶斯预测的均方误差($n = 100$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$1\,000M_{SE_B}$	0.828 5	0.803 5	0.933 3	0.836 9	0.838 5	0.830 4	0.814 1	0.900 3
$1\,000M_{SE_C}$	0.561 2	0.536 9	0.613 4	0.563 3	0.560 6	0.576 1	0.538 5	0.590 8
$1\,000M_{SE_D}$	0.042 8	0.043 4	0.042 8	0.042 9	0.042 9	0.042 7	0.043 5	0.043 2

表 9 在二项-贝塔分布模型中贝叶斯预测的均方误差($n = 200$)

均方误差	保单数							
	1	2	3	4	5	6	7	8
$1\,000M_{SE_B}$	0.201 6	0.219 6	0.233 5	0.206 8	0.225 3	0.205 5	0.221 7	0.232 7
$1\,000M_{SE_C}$	0.166 4	0.176 4	0.190 3	0.165 2	0.185 7	0.166 6	0.179 5	0.187 9
$1\,000M_{SE_D}$	0.021 5	0.021 6	0.021 0	0.021 8	0.021 1	0.021 1	0.021 2	0.021 3

根据表 7 ~ 表 9 可知,在二项-贝塔模型中均方误差的大小关系为 $M_{SE_D} < M_{SE_C} < M_{SE_B}$. 即此时信度预测 $\hat{M}_i^D$ 有最小的均方误差. 而其他 2 个预测由于对模型的误识别而导致均方误差比信度预测 $\hat{M}_i^D$ 更大一些. 这种差异在小样本($n = 10$)时比较明显,随着样本容量增大,3 个预测的均方误差的差值有减少的趋势.

4 小结

本文给出了在 3 种模型下未来索赔次数的贝叶斯预测和信度预测,得到了均方预测误差的表达式,并讨论了 3 个预测的统计性质. 在非寿险特别是汽车保险中,保单的索赔次数常常服从泊松分布. 然而,由于风险的异质性,所以风险参数的先验分布往往是未知的. 这时若对先验分布的分布类型进行假设,则可能会因为模型的误识别而导致得到错误的预测,使得预测值与真实值出现较大的误差. 这时,信度预测 $\hat{M}_i^C$ 显然是一个好的选择. 当然,若非泊松分布或者样本分布未知,则建议使用 $\hat{M}_i^D$ 作为未来索赔次数的预测,能更加准确地预测未来的索赔次数,进而制定合适的奖惩系统或为保单制定“差异费率”.

5　参考文献

[1] 温利民. 信度估计的理论与方法 [M]. 北京: 科学出版社,2012.

[2] 张连增,申晴. 泊松提升模型在中国车险索赔频率预测建模中的应用 [J]. 统计与信息论坛,2019,34(9): 27-34.

[3] 王选鹤,孟生旺,杨默. 车险索赔次数预测模型的扩展与应用 [J]. 保险研究,2018(11): 82-92.

[4] 孟生旺,杨亮. 随机效应零膨胀索赔次数回归模型 [J]. 统计研究,2015,32(11): 97-102.

[5] Li Bo, Ni Weihong, Constantinescu C. Risk models with premiums adjusted to claims number [J]. Insurance: Mathematics and Economics,2015,65: 94-102.

[6] Dembińska A, Buraczyńska A. The long-term behavior of number of near-maximum insurance claims [J]. Insurance: Mathematics and Economics,2019,88: 226-237.

[7] Li Yun, Pakes A G. On the number of near-maximum insurance claims [J]. Insurance: Mathematics and Economics,2001,28(3): 309-323.

[8] Denuit M, Maréchal X, Pitrebois S, et al. Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems [M]. Chichester: John Wiley and Sons,2007: 119-163.

[9] Young V R. Credibility using semiparametric models and a loss function with a constancy penalty [J]. Insurance: Mathematics and Economics,2000,26(2/3): 151-156.

[10] Canfield R V. A Bayesian approach to reliability estimation using a loss function [J]. IEEE Transactions on Reliability,1970,19(1): 13-16.

[11] Gómez-Déniz E. A generalization of the credibility theory obtained by using the weighted balanced loss function [J]. Insurance: Mathematics and Economics, 2008, 42(2): 850-854.

[12] 茆诗松,王静龙,濮晓龙. 高等数理统计 [M]. 北京: 高等教育出版社,2006.

[13] Morata L B I. A priori ratemaking using bivariate Poisson regression models [J]. Insurance: Mathematics and Economics,2009,44(1): 135-141.

[14] Bermúdez L, Karlis D. A posteriori ratemaking using bivariate Poisson models [J]. Scandinavian Actuarial Journal, 2017(2): 148-158.

[15] Bermúdez L, Karlis D. Bayesian multivariate Poisson models for insurance ratemaking [J]. Insurance: Mathematics and Economics,2011,48(2): 226-236.

[16] Berkhout P, Plug E. A bivariate poisson count data model using conditional probabilities [J]. Statistica Neerlandica, 2004,58(3): 349-364.

[17] Bühlmann H, Gisler A. A course in credibility theory and its applications [M]. Berlin: Springer-Verlag, 2005: 55-75.

[18] 程兵,陈萍. 具有风险相依结构的平衡信度估计 [J]. 经济数学,2016,33(1): 80-83.

[19] 章溢,李志龙,温利民. 矩相关保费原理中风险保费的经验厘定 [J]. 中国科学: 数学,2019,49(7): 1041-1062.

[20] 章溢,熊佳,温利民,等. 基于核估计下概率密度函数的信度模型 [J]. 高校应用数学学报,2020,35(1): 29-39.

[21] Gisler A, Wüthrich M V. Credibility for the chain ladder reserving method [J]. Astin Bulletin,2008,38(2): 565-600.

[22] Kremer E. A remark on parameter-estimation of autoregressive credibility-models [J]. Blätter der DGVFM, 1983, 16(2): 153-159.

[23] Atanasiu V. Applications aiming Bühlmann´s credibility model [J]. University Politehnica of Bucharest Scientific Bulletin: Seriers A,2011,73(2): 51-64.

The Bayesian Prediction and Credibility Approximate of Claim Number

ZHANG Yi[1], ZHOU Jinliang[2]

(1. School of Finance, Jiangxi Normal University, Nanchang Jiangxi 330022, China;
2. School of Mathematics and Statistics, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: The Bayesian model is established for the number of claims of the insurance policy. According to whether the sample distribution and the prior distribution are known, the Bayesian prediction and credibility approximate claim number are discussed in three kinds of situations, and the estimation method of structural parameters is given. Finally, the convergences and the mean square errors of the three predictions are compared by numerical simulation.

Key words: Poisson distribution; Gamma distribution; Bayesian prediction; credibility approximation; structural parameters

(责任编辑: 曾剑锋)