续表 2

θ_0	a	γ											
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	γ_s
0.8	0.025	0.566	0.641	0.686	0.732	0.754	0.773	0.743	0.663	0.488	0.26	0.107	0.669
	0.05	0.713	0.782	0.828	0.874	0.904	0.919	0.884	0.802	0.595	0.331	0.134	0.811
	0.075	0.79	0.845	0.894	0.932	0.956	0.961	0.947	0.874	0.664	0.356	0.135	0.863
0.9	0.025	0.329	0.424	0.523	0.638	0.69	0.719	0.694	0.606	0.421	0.238	0.118	0.639
	0.05	0.449	0.548	0.658	0.767	0.831	0.867	0.845	0.748	0.549	0.333	0.186	0.781
	0.075	0.512	0.64	0.75	0.838	0.89	0.935	0.931	0.82	0.649	0.428	0.25	0.841

3.3　实证分析

消费者价格指数(consumer price index,CPI)用来表示通货膨胀的程度,是反映一国经济发展情况的一个非常重要的经济指标.将美国 1970 年至 1990 年共 21 年的平均 CPI[14]转换为环比增长率数据,如图 5 所示.在不同的调节参数取值的情况下,应用 CUSUM 方法估计该数据序列的均值变点位置,结果如表 3 所示.由表 3 的估计结果可以看到,调节参数 γ 的选取对变点估计的结果会有影响,在调节参数 γ 取值较小时,CUSUM 方法估计得到的变点位置为 1981 年,而在调节参数 γ 取值较大时,估计得到的变点位置为 1982 年.

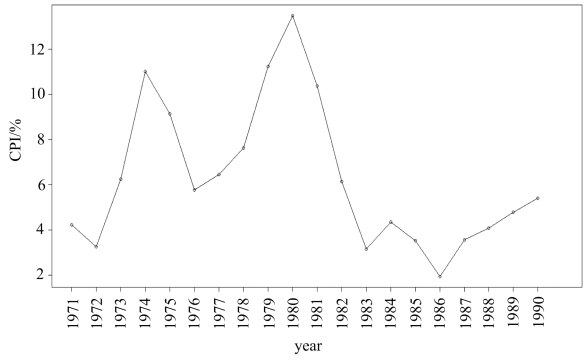

图 5　美国 1971 年至 1990 年 CPI 环比增长率数据

Fig. 5　Year-to-year growth rate of US annual average CPI from 1971 to 1990

表 3　不同调节参数情况下的变点估计结果

Tab. 3　Change-point estimations under different values of tuning parameter

year	$\hat{k}$	γ
1981	11	0
1981	11	0.1
1981	11	0.2
1982	12	0.3
1982	12	0.4
1982	12	0.5
1982	12	0.6
1982	12	0.7
1982	12	0.8
1982	12	0.9

为检验本文提出的基于数据驱动的调节参数选择算法在实际应用中的有效性,将其应用到上述美国 CPI 变点估计问题中.令调节参数 γ 的可能取值集为 $\gamma=\left\{\frac{j}{200}, j=1,2,\cdots,200\right\}$,应用所提出的基于数据驱动的调节参数 γ 的选取方法,选定 1971 年至 1990 年美国年度平均 CPI 环比增长率数据所适用的调节参数 $\gamma_s=0.24$,并计算 CUSUM 型变点估计量(2),得出变点估计位置为 1982 年.

王伟[15]利用基于 M 估计的线性回归模型检验 1970 年至 1990 年美国年度平均 CPI 的变点时,将 1981 年作为变点.我们应用 CUSUM 方法给出美国 CPI 的变点估计值为 1982 年,更符合实际.因为美国在 20 世纪 70 年代通货膨胀严重,出现了经济萧条的情况,1980 年至 1982 年期间经历了第二次经济危机,后来经过一系列的宏观调控,1982 年结束了经济危机,扭转了通货膨胀的态势,1983 年恢复了经济增长,所以我们认为 1982 年为美国消费者价格指数的变点更合适.

4　结论

CUSUM 方法作为变点检测和估计最常用的非参数方法之一,其调节参数取值的选择一直备受关注.本文分别对独立样本和相依样本序列,通过蒙特卡罗模拟方法探究了调节参数取值对 CUSUM 型变点估计量的估计效果的影响,并提出了一种基于数据驱动的调节参数选取法,证明了该方法的稳健性,这为带有调节参数的 CUSUM 方法在变点问题统计推断中的应用提供了理论支撑.实证分析显示,基于数据驱动方法给出的调节参数取值在实际数据分析中具有明显的应用价值.但是模拟结果同时发现,基于数据驱动方法给出的调节参数取值并不是最优的,如何基于数据驱动方法给出一个最优的调节参数取值,这将是后续需要解决的问题.

参考文献(References)

[1] PAGE E S. Continuous inspection schemes [J]. Biometrika, 1954, 41: 100-115.

[2] GIRAITIS L, LEIPUS R, SURGAILIS D. The change-point problem for dependent observations[J]. Statistical Planning and Inference, 1996, 53(3): 297-310.

[3] CSORGO M, HORWLTH L. Limit Theorems in Change-Point Analysis [M]. Chichester, UK: John Wiley and Sons Ltd, 1997.

[4] HORVATH L, KOKOSZKA P. The effect of long-range dependence on change-point estimators [J]. Statistical Planning and Inference, 1997, 64(1):

57-81.
[5] KOKOSZKA P, LEIPUS R. Testing for parameter changes in ARCH models [J]. Lithuanian Mathematical Journal, 1999, 39(2): 182-195.
[6] HARIZ S B, WYLIE J J. Rates of convergence for the change-point estimator for long-range dependent sequences[J]. Statistics and Probability Letters, 2005, 73(2): 155-164.
[7] HARIZ S B , WYLIE J J, ZHANG Q. Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences[J]. The Annals of Statistics, 2007, 35(4): 1802-1826.
[8] NIE W L , HARIZ S B, WYLIE J J, et al. Change-point detection for long-range dependent sequences in a general setting[J]. Nonlinear Analysis, 2009, 71(12): 2398-2405.
[9] FREMDT S. Asymptotic distribution of the delay time in Page's sequential procedure[J]. Statistical Planning and Inference, 2013, 145: 74-91.
[10] CHEN Z H, HU Y J. Cumulative sum estimator for change-point in panel data [J]. Statistical Papers, 2017, 58: 707-728.
[11] QIN R B, MA J J. An efficient algorithm to estimate the change in variance[J]. Economics Letters, 2018, 168:15-17.
[12] XU M, WU Y, JIN B. Detection of a change-point in variance by a weighted sum of powers of variances test [J]. Journal of Applied Statistics, 2019, 46(4): 664-679.
[13] TAN C C, SHI X P, WU Y H. On nonparametric change point estimator based on empirical characteristic functions[J]. Science China Mathematics, 2016, 59(12): 2463-2484.
[14] Federal Reserve Bank of St. Louis. Consumer price index for all urban consumers: All items in U. S. city average (CPIAUCSL) [EB/OL]. [2020-04-01]. https://fred. stlouisfed. org/series/CPIAUCSL.
[15] 王伟. 基于 M 估计的线性回归模型均值变点检测[D]. 南京:东南大学, 2011.

第50卷第7期 中国科学技术大学学报 Vol. 50, No. 7
2020年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2020
文章编号:0253-2778(2020)07-0927-07

保险失效或退保概率的马尔可夫链预测模型

张 影

(中国科学技术大学数学科学学院,安徽合肥 230026)

摘要:利用连续时间时齐马尔可夫链构建关于保险失效或退保概率的预测模型,用以计算任一时刻处于各个状态的概率,并给出参数估计的方法.由于实际中保单的状态在特定的时刻会发生离散事件,因此使用多阶段的马尔可夫链模型来刻画这一特点,即在发生离散事件的特定时刻,定义一个跳跃矩阵描述该时刻的状态转移情况.将该模型应用到保险公司的寿险失效或退保概率的研究中,通过实际的数据对模型参数进行估计和校准,通过校准的马尔可夫链模型对其寿险失效或退保概率进行预测并得到较好的预测结果.

关键词:马尔可夫链;强度矩阵;转移概率;寿险失效或退保概率预测
中图分类号:O211.62 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.07.009
2020 Mathematics Subject Classification:60J28

引用格式:张影. 保险失效或退保概率的马尔可夫链预测模型[J]. 中国科学技术大学学报,2020,50(7):929-935.
ZHANG Ying. Markov chains prediction model of insurance lapse or surrender probability[J]. Journal of University of Science and Technology of China, 2020,50(7):929-935.

Markov chains prediction model of insurance lapse or surrender probability

ZHANG Ying

(*School of Mathematics Sciences, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: The continuous-time time-homogeneous Markov chain was used to construct a prediction model about the probability of insurance lapse or surrender to calculate the probability of being in various states at any time, and a parameter estimation method was given. In reality, the state of the insurance policy would have discrete events at a specific time, so the multi-stage Markov chain model was used to characterize this feature. That was, at a specific time when a discrete event occurs, a jump matrix was defined to describe the state transition at the specific time. The model was applied to the study of the life insurance lapse or surrender probability of insurance companies, and the model parameters were estimated and calibrated by actual data. Finally, the calibrated Markov chain model was used to predict the life insurance lapse or surrender probability and a good prediction result was obtained.

Key words: Markov chains; intensity matrix; transition probability; life insurance lapse or surrender probability prediction

0 引言

退保行为主要发生在寿险行业中,投保人在保险期内可申请退保.孙蓉等[1]指出退保行为不仅会降低保险公司的利润,如果退保金额过大,还可能会带来现金流中断的危险.因此保险公司一直对退保行为感到担忧,而险种的退保概率则直接反映了保险公司退保行为发生的严重程度,故对退保概率的建模和预测一直是保险公司研究的主要课题之一.

目前,对退保问题的研究主要集中在退保行为及其影响因素上面,文献[2-5]分析了一些宏观与微观因素对退保行为的影响.Milhaud[6]利用外源性和内源性的风险因素对退保行为进行建模.毕泗锋和亓超[7]从人的认知偏差角度来研究影响退保行为的因素.王向楠[8]利用计数模型和负二项回归从产品结构角度实证研究中国寿险产品的退保问题.郭春燕[9]利用Cox比例危险模型,根据不同的数据特点,对退保率建立了不同的预测模型.Kim[10]通过利率与失业率等影响因素,利用Logit和CLL模型来研究投保人的退保行为.Neves等[11]则提出了多阶段随机模型,利用投保人变量和金融变量来预测退保率.

我们注意到寿险涉及多状态转移的问题,即投

收稿日期:2020-04-29;**修回日期**:2020-06-10
作者简介:张影,女,1994年生,硕士.研究方向:随机过程及其应用.E-mail:zy028023@mail.ustc.edu.cn

保人在投保之后保单状态可以随着时间转换成其他状态,因此,本文利用时齐马尔可夫链(简称马氏链)来构建关于失效或退保概率的预测模型,并且在发生离散事件的时刻定义一个跳跃矩阵来描述状态转移情况.通过保险公司实际数据对模型常数强度矩阵和跳跃矩阵进行参数估计与校准,进而可得到险种各状态分布的概率预测值.最后将我国某寿险公司的六种险种的数据代入模型中,在固定样本的前提下,可以得到保单在投保后每一年的失效或退保概率,结果表明此模型有较好的预测效果.

1 马尔可夫链相关理论基础

设 $(\Omega,\mathscr{F},\mathbb{P})$ 为一完备的概率空间.在该概率空间下,设 $M=\{M_t\}_{t\geqslant 0}$ 为一连续时间有限状态马氏链.不失一般性,设其状态空间为 $\mathscr{H}=\{1,2,\cdots,H\}$,$H$ 为马氏链的状态个数.设 $\mathscr{F}_t=\sigma(M_s;s\in[0,t])$,$t\geqslant 0$,为由马氏链 M 生成的自然 σ 代数流,于是当 $s<t$ 时,$\mathscr{F}_s\subseteq\mathscr{F}_t$.现在,假设 $M=\{M_t\}_{t\geqslant 0}$ 为时齐的马氏链,这意味着对 $\forall s,t\geqslant 0$,其转移概率函数为

$$p_{ij}(t)\triangleq\mathbb{P}(M_{t+s}=j\mid M_s=i),\forall i,j\in\mathscr{H}.$$

则马氏链 $M=\{M_t\}_{t\geqslant 0}$ 的转移概率矩阵族 $P(t)=[p_{ij}(t)]_{1\leqslant i,j\leqslant H}$,$t\geqslant 0$,此处 $P(0)=Id=[\delta_{ij}]_{1\leqslant i,j\leqslant H}$,其中 $\delta_{ij}=\mathbf{1}_{\{i=j\}}$.进一步,文献[12]可知下面的极限存在:

$$q_{ij}\triangleq\lim_{t\downarrow 0}\frac{p_{ij}(t)-p_{ij}(0)}{t}=\lim_{t\downarrow 0}\frac{p_{ij}(t)-\delta_{ij}}{t},\quad\forall i,j\in\mathscr{H}.$$

称 $Q=[q_{ij}]_{1\leqslant i,j\leqslant H}$ 为马氏链 M 的强度矩阵.注意到由于马氏链 M 为时间齐次的,那么其强度矩阵 Q 为常数矩阵.文献[13]指出时齐马氏链 M 的转移概率矩阵 $P(t)$ 与强度矩阵 Q 的关系为

$$P(t)=\mathrm{e}^{tQ}=\sum_{n=0}^{\infty}\frac{Q^n t^n}{n!},\forall t\geqslant 0.$$

设 H 维行向量

$$\vec{\mu}_0=[\mu_0(i)]_{1\leqslant i\leqslant H}=[\mathbb{P}(M_0=i)]_{1\leqslant i\leqslant H}$$

表示马氏链 M 在概率测度 $\mathbb{P}$ 下的初始概率分布.同样地,对任意 $t\geqslant 0$,用向量 $\vec{\mu}_t=[\mu_t(i)]_{1\leqslant i\leqslant H}=[\mathbb{P}(M_t=i)]_{1\leqslant i\leqslant H}$ 表示马氏链 M 在概率测度 $\mathscr{P}$ 下的 t 时刻概率分布.则对 $\forall t\geqslant 0$ 有

$$\vec{\mu}_t=\vec{\mu}_0P(t)=\vec{\mu}_0\mathrm{e}^{tQ}.$$

2 基于马尔可夫链的多状态转移数学模型及其求解

2.1 变量定义

设 $\mathscr{H}=\{1,2,\cdots,H\}$ 是保单的状态集合.$\tau_i^n(n\in\mathbb{N})$ 是指保单在发生第 $n-1$ 次状态转移时到达状态 i 后停留在状态 i 的时长.$\eta_i^n(n\in\mathbb{N})$ 取值于 $\mathscr{H}\backslash\{i\}$,是指保单在第 $n-1$ 次状态转移时到达状态 i,发生第 n 次状态转移时,保单由状态 i 转移出去后所到达的状态.随机过程 $X=\{X_t\}_{t\in[0,T]}$ 取值于 $\mathscr{H}$,代表保单在时间 $[0,T]$ 内的状态转移过程.设 $\vec{\mu}_t=[\mathbb{P}(X_t=i)]_{1\leqslant i\leqslant H}$ 代表随机过程 $X=\{X_t\}_{t\in[0,T]}$ 在时刻 $t\in[0,T]$ 时的概率分布.

2.2 模型假设

现在对模型进行如下假设:

(A1) 保单在时间 $[0,T)$ 内任意时刻都可能发生状态转移,即随机过程 X 在时间 $[0,T)$ 内是连续时间状态转移过程.如果保单在 $[0,T)$ 内发生状态转化,则只能转化成 $\{1,2,\cdots,m\}(m\in\mathbb{N},1\leqslant m\leqslant H)$ 中的任意状态,随机过程 $X=\{X_t\}_{t\in[0,T]}$ 的分布 $\vec{\mu}_t$ 关于 $t\in[0,T)$ 连续.

(A2) 保单在时刻 T 时发生离散状态转移事件.保单在时刻 T 可以转化成 $\{1,2,\cdots,m,m+1,\cdots,H\}$ 中的任意状态,随机过程 $X=\{X_t\}_{t\in[0,T]}$ 的分布 $\vec{\mu}_t$ 在 $t=T$ 处不连续,发生跳跃.

(A3) 随机过程 $X=\{X_t\}_{t\in[0,T]}$ 在 $t\in[0,T)$ 是连续时间时齐马氏链,且强度矩阵 $Q=[q_{ij}]_{1\leqslant i,j\leqslant H}$ 存在.

2.3 模型分析

由于 T 时刻发生的离散事件导致分布 $\vec{\mu}_t$ 在 $t=T$ 处发生跳跃,即 $\vec{\mu}_{T-}$ 向 $\vec{\mu}_T$ 跳跃了一下,因此本文从时间 $[0,T)$ 内发生状态转移和时刻 T 时发生状态转移两部分来对模型进行分析.

2.3.1 $[0,T)$内状态转移的连续过程分析

由文献[14]对连续时间时齐马氏链的构造可知保单的停留时长和转出状态满足下面三个条件:

(a)对每个给定的状态 $i\in\mathscr{H}$,停留时长 $\tau_i^n(n\in\mathbb{N})$ 同分布于参数为 $r_i=-q_{ii}$ 的指数分布.

(b)对每个给定的状态 $i\in\mathscr{H}$,转出状态 $\eta_i^n(n\in\mathbb{N})$ 取值于 $\mathscr{H}\backslash\{i\}$,并且同分布,它的概率分布为

$$\mathbb{P}(\eta_i^n=j)=-\frac{q_{ij}}{q_{ii}},\forall j\neq i.$$

(c)随机变量 $\{\tau_i^n,\eta_i^n\}_{i\in\mathscr{H},n\in\mathbf{N}}$ 相互独立.

进一步,由模型假设(A1)和假设(A3)可知:若随机过程 $X=\{X_t\}_{t\in[0,T]}$ 的初始状态 X_0 的分布为 $\vec{\mu}_0$,而随机过程 $X=\{X_t\}_{t\in[0,T]}$ 在 $t\in[0,T)$ 是连续时间的时齐马氏链,则对 $\forall t\in[0,T)$,随机过程 $X=\{X_t\}_{t\in[0,T]}$ 的分布 $\vec{\mu}_t$ 为

$$\vec{\mu}_t=\vec{\mu}_0\mathrm{e}^{tQ}.$$

2.3.2 T 时刻的状态转移情况分析

由假设(A2),随机过程 $X=\{X_t\}_{t\in[0,T]}$ 的分布 $\vec{\mu}_t$ 在 $t=T$ 处不连续,发生跳跃,不妨定义 $\vec{\mu}_{T-}$ 到 $\vec{\mu}_T$ 之间的变换矩阵为跳跃矩阵 J,即 $\vec{\mu}_T=\vec{\mu}_{T-}J$.由于 $\vec{\mu}_t=\vec{\mu}_0\mathrm{e}^{tQ}$ 关于 $t\in[0,T)$ 是连续函数,于是 $\vec{\mu}_{T-}=\vec{\mu}_0\mathrm{e}^{(T-)Q}=\vec{\mu}_0\mathrm{e}^{TQ}$.进一步,可以得到随机过程 $X=\{X_t\}_{t\in[0,T]}$ 在 T 时刻的分布 $\vec{\mu}_T$ 为

$$\vec{\mu}_T=\vec{\mu}_0\mathrm{e}^{TQ}J.$$

2.4 模型求解

2.4.1 强度矩阵 Q 的对角元素的估计

假设对样本进行 $n(n\in\mathbb{N})$ 次独立随机抽样,在第 $k(k\in\mathbb{N},k\leqslant n)$ 次抽样中,对每个给定的状态 $i\in\mathscr{H}$,有 $m^k(i)$ 次到达状态 i,停留的时间依次

为 $t_1^k(i),t_2^k(i),\cdots,t_{m^k(i)}^k(i)$，我们定义统计量 $T_{ii}(n)$ 为

$$T_{ii}(n):=\frac{\sum_{k=1}^{n}(\sum_{\alpha=1}^{m^k(i)}t_\alpha^k(i))}{\sum_{k=1}^{n}m^k(i)},\ i\in\mathscr{H},\ n\in\mathbb{N}.$$

于是有如下主要结果：

定理 2.1　假设对每个给定的状态 $i\in\mathscr{H}$，停留时长 $\tau_i^n(n\in\mathbb{N})$ 满足模型分析部分的条件(a)和条件(c)，那么连续时间时齐马氏链的强度矩阵 Q 的对角元素 q_{ii} 的统计量为

$$\hat{q}_{ii}(n)=-\frac{1}{T_{ii}(n)}=-\frac{\sum_{k=1}^{n}m^k(i)}{\sum_{k=1}^{n}(\sum_{\alpha=1}^{m^k(i)}t_\alpha^k(i))},\quad i\in\mathscr{H},\ n\in\mathbb{N}.$$

并且有

$$E[\hat{q}_{ii}(n)]\to q_{ii},\ n\to\infty,$$

即，$\hat{q}_{ii}(n)$ 是 q_{ii} 的渐进无偏估计.

证明　由模型分析部分的条件(a)、条件(c)和独立随机抽样可知，对于给定的状态 $i\in\mathscr{H}$ 和任意的整数 $k,\alpha(1\leqslant k\leqslant n,\ 1\leqslant\alpha\leqslant m^k(i))$，$t_\alpha^k(i)$ 独立同分布于参数为 $-q_{ii}$ 的指数分布，即随机变量 $t_\alpha^k(i)$ 的概率密度函数为

$$f(x)=\begin{cases}-q_{ii}e^{q_{ii}x},\ x\geqslant 0;\\ 0,\ x<0.\end{cases}$$

则随机变量 $t_\alpha^k(i)$ 的数学期望为

$$E[t_\alpha^k(i)]=-\frac{1}{q_{ii}},\forall\, i\in\mathscr{H}.$$

即可知对 $\forall\, n\geqslant 1$，$T_{ii}(n)$ 的数学期望为

$$E[T_{ii}(n)]=\frac{\sum_{k=1}^{n}(\sum_{\alpha=1}^{m^k(i)}E[t_\alpha^k(i)])}{\sum_{k=1}^{n}m^k(i)}=\frac{\sum_{k=1}^{n}(-m^k(i)\frac{1}{q_{ii}})}{\sum_{k=1}^{n}m^k(i)}=-\frac{1}{q_{ii}},$$

即得 $T_{ii}(n)$ 是 $-\frac{1}{q_{ii}}$ 的无偏估计. 再由强大数定律可知 $\mathbb{P}$-a. s.，

$$T_{ii}(n)\to-\frac{1}{q_{ii}},\ n\to\infty.$$

即在概率测度 $\mathbb{P}$，$T_{ii}(n)$ 几乎处处收敛到 $-1/q_{ii}$.

现在定义函数 $g(x):=-\frac{1}{x}$，其中 $x>0$，则 g 在 $(0,+\infty)$ 上连续. 那么应用连续映射定理可知 $\mathbb{P}$-a. s.，

$$g(T_{ii}(n))\to g(-\frac{1}{q_{ii}}),\ n\to\infty,$$

也就是 $\mathbb{P}$-a. s.，

$$-\frac{1}{T_{ii}(n)}\to q_{ii},\ n\to\infty.$$

进一步，应用有界收敛定理，得到

$$E[-\frac{1}{T_{ii}(n)}]\to q_{ii},\ n\to\infty.$$

即 $\hat{q}_{ii}(n)$ 是 q_{ii} 的渐进无偏估计.

2.4.2　强度矩阵 Q 的非对角元素的估计

假设对样本进行 $n(n\in\mathbb{N})$ 次独立随机抽样，在第 $k(k\in\mathbb{N},\ k\leqslant n)$ 次抽样中，对每个给定的状态 $i\in\mathscr{H}$，有 $m^k(i)$ 次到达状态 i，停留完成后进行转移，其中有 $m^k(i,j)$ 次由 i 状态转移至 $j(j\in\mathscr{H},\ j\neq i)$ 状态. 定义统计量 $T_{ij}(n)$ 为如下形式：

$$T_{ij}(n)=\frac{\sum_{k=1}^{n}m^k(i,j)}{\sum_{k=1}^{n}m^k(i)},\ i,j\in\mathscr{H},\ i\neq j,\ n\in\mathbb{N}.$$

于是有如下定理：

定理 2.2　假设对每个给定的状态 $i\in\mathscr{H}$，停留时长 $\tau_i^n(n\in\mathbb{N})$ 和转出状态 $\eta_i^n(n\in\mathbb{N})$ 满足模型分析部分的条件(a)～(c)，那么连续时间时齐马氏链的强度矩阵 Q 的非对角元素 $q_{ij}(j\in\mathscr{H},\ i\neq j)$ 的统计量为

$$\begin{cases}\hat{q}_{ij}(n)=\dfrac{T_{ij}(n)}{T_{ii}(n)}=\dfrac{\sum_{k=1}^{n}m^k(i,j)}{\sum_{k=1}^{n}(\sum_{\alpha=1}^{m^k(i)}t_\alpha^k(i))},\\ i,j\in\mathscr{H},\ i\neq j,\ n\in\mathbb{N}\end{cases}\tag{1}$$

进一步，我们有

$$E[\hat{q}_{ij}(n)]\to q_{ij},\ n\to\infty,$$

即，$\hat{q}_{ij}(n)$ 是 q_{ij} 的渐进无偏估计.

证明　由模型分析部分的条件(a)～(c)和独立随机抽样知：随机变量 $m^k(i,j)$ 是 $m^k(i)$ 个独立同分布的伯努利分布随机变量的和，即随机变量 $m^k(i,j)$ 的概率分布为

$$\mathbb{P}(m^k(i,j)=m)=\binom{m^k(i)}{m}\left(-\frac{q_{ij}}{q_{ii}}\right)^m\left(1+\frac{q_{ij}}{q_{ii}}\right)^{m^k(i)-m}.$$

则可得 $T_{ij}(n)$ 的数学期望为

$$E[T_{ij}(n)]=\frac{\sum_{k=1}^{n}E[m^k(i,j)]}{\sum_{k=1}^{n}m^k(i)}=\frac{\sum_{k=1}^{n}(-m^k(i)\cdot\frac{q_{ij}}{q_{ii}})}{\sum_{k=1}^{n}m^k(i)}=-\frac{q_{ij}}{q_{ii}}.$$

即 $T_{ij}(n)$ 是 $-\frac{q_{ij}}{q_{ii}}$ 的无偏估计. 再由强大数定律可

知 $\mathbb{P}$-a.s.，

$$T_{ij}(n) \to -\frac{q_{ij}}{q_{ii}},\ n \to \infty.$$

因为前面已经得到 $-\frac{1}{T_{ii}(n)} \to q_{ii}$，$n \to \infty$，则有 $\mathbb{P}$-a.s.，

$$\left(-\frac{1}{T_{ii}(n)}, T_{ij}(n)\right) \to \left(q_{ii}, -\frac{q_{ij}}{q_{ii}}\right),\ n \to \infty.$$

现在定义二元函数 $g(x,y):=-xy$，其中 x，$y \in \mathbb{R}$，则函数 g 在 $\mathbb{R}^2$ 上连续. 因此应用连续映射定理得 $\mathbb{P}$-a.s.，

$$g\left(-\frac{1}{T_{ii}(n)}, T_{ij}(n)\right) \to g\left(q_{ii}, -\frac{q_{ij}}{q_{ii}}\right),\ n \to \infty,$$

即 $\mathbb{P}$-a.s.，

$$\frac{T_{ij}(n)}{T_{ii}(n)} \to q_{ij},\ n \to \infty.$$

再应用有界收敛定理得

$$E\left[\frac{T_{ij}(n)}{T_{ii}(n)}\right] \to q_{ij},\ n \to \infty.$$

即 $\hat{q}_{ij}(n)$ 是 q_{ij} 的渐进无偏估计.

2.4.3 跳跃矩阵 J 的估计

假设对样本进行 $n(n \in \mathbb{N})$ 次独立随机抽样，对每个给定的状态 $i \in \mathscr{H}$，在 T 时刻状态为 i 的样本数为 $n(i)$，在 $n(i)$ 个样本中，在 T 时刻由状态 i 转移至状态 $k \in \mathscr{H}$ 的样本数为 $n(i,k)$，于是有如下定理：

定理 2.3 假设对每个给定的状态 $i \in \mathscr{H}$，在 T 时刻状态分布的变换矩阵为跳跃矩阵 J，那么跳跃矩阵 J 的元素 $j_{ik}(k \in \mathscr{H})$ 的统计量 $\hat{j}_{ik}(n)$ 为

$$\hat{j}_{ik}(n) = \frac{n(i,k)}{n(i)},\ i,k \in \mathscr{H},\ n \in \mathbb{N} \tag{2}$$

进一步，有

$$E[\hat{j}_{ik}(n)] = j_{ik},$$

即，$\hat{j}_{ik}(n)$ 是 j_{ik} 的无偏估计.

证明 由于保单状态在 T 时刻的转移概率矩阵为 J，再由模型假设和随机抽样可知 $n(i,k)$ 是 $n(i)$ 个独立同分布的伯努利分布随机变量的和，即随机变量 $n(i,k)$ 的概率分布为

$$\mathbb{P}(n(i,k)=m) = \binom{n(i)}{m} j_{ik}^m (1-j_{ik})^{n(i)-m}.$$

因此可得 $\hat{j}_{ik}(n)$ 的数学期望为

$$E[\hat{j}_{ik}(n)] = \frac{E[n(i,k)]}{n(i)} = \frac{n(i)\cdot j_{ik}}{n(i)} = j_{ik}.$$

由此可知 $\hat{j}_{ik}(n)$ 是 j_{ik} 的无偏估计. 再由强大数定律可知 $\mathbb{P}$-a.s.，

$$\hat{j}_{ik}(n) \to j_{ik},\ n \to \infty.$$

得证.

2.5 马氏链模型的概率分布预测

由前面建立的连续时间时齐马氏链数学模型可知：给定初始分布 $\vec{\mu}_0$，对任意 $t \in [0,T]$，随机过程 $X=\{X_t\}_{t\in[0,T]}$ 在 t 时刻的分布 $\vec{\mu}_t$ 为

$$\vec{\mu}_t = \begin{cases} \vec{\mu}_0 e^{tQ}, & 0 \leqslant t < T; \\ \vec{\mu}_0 e^{TQ} J, & t = T. \end{cases}$$

通过对样本进行 $n(n \in \mathbb{N})$ 次独立随机抽样以及前面的参数估计，可以得到强度矩阵 Q 的参数估计 $\hat{Q}(n)=[\hat{q}_{ik}(n)]_{1\leqslant i,k\leqslant H}$，也可以得到跳跃矩阵 J 的参数估计 $\hat{J}(n)=[\hat{j}_{ik}(n)]_{1\leqslant i,k\leqslant H}$. 将 $\hat{Q}(n)$ 和 $\hat{J}(n)$ 代入上式就能得到随机过程 $X=\{X_t\}_{t\in[0,T]}$ 在 $t \in [0,T]$ 时的预测分布 $\hat{\vec{\mu}}_t(n)$ 为

$$\hat{\vec{\mu}}_t(n) = \begin{cases} \vec{\mu}_0 e^{t\hat{Q}(n)}, & 0 \leqslant t < T; \\ \vec{\mu}_0 e^{T\hat{Q}(n)} \hat{J}(n), & t = T \end{cases} \tag{3}$$

因此有如下定理：

定理 2.4 假设对每个给定的状态 $i \in \mathscr{H}$，停留时长 $\tau_i^n(n \in \mathbb{N})$ 和转出状态 $\eta_i^n(n \in \mathbb{N})$ 满足模型分析部分的条件(a)～(c)，那么对 $\forall t \in [0,T]$，式(3)中的 $\hat{\vec{\mu}}_t(n)$ 即为分布 $\vec{\mu}_t$ 的统计量，并且有

$$\hat{\vec{\mu}}_t(n) \to \vec{\mu}_t,\ \forall t \in [0,T],\ n \to \infty,$$

即，$\hat{\vec{\mu}}_t(n)$ 为 $\vec{\mu}_t$ 的渐进无偏估计.

证明 由定理 2.1 和定理 2.2 可知参数估计 $\hat{Q}(n)$ 是强度矩阵 Q 的渐进无偏估计. 现在定义函数 $g(x) \triangleq \vec{\mu}_0 e^{tx}$，其中 $x \in \mathbb{R}^{H\times H}$，则 g 在 $\mathbb{R}^{H\times H}$ 上连续. 应用连续映射定理可得 $\mathbb{P}$-a.s.，

$$g(\hat{Q}(n)) \to g(Q),\ n \to \infty,$$

即 $\mathbb{P}$-a.s.，

$$\hat{\vec{\mu}}_t(n) = \vec{\mu}_0 e^{t\hat{Q}(n)} \to \vec{\mu}_0 e^{tQ},\ \forall t \in [0,T),$$

即得在 $t \in [0,T)$ 时，$\hat{\vec{\mu}}_t(n)$ 是 $\vec{\mu}_t$ 的渐进无偏估计. 进一步，应用定理 2.3 得 $\mathbb{P}$-a.s.，

$$\hat{J}(n) \to J,\ n \to \infty.$$

因此有 $\mathbb{P}$-a.s.，

$$(\hat{\vec{\mu}}_t(n), \hat{J}(n)) \to (\vec{\mu}_t, J),\ n \to \infty.$$

现在定义函数 $l(x,y):=\vec{\mu}_0 e^{tx} y$，其中 $x,y \in \mathbb{R}^{H\times H}$，则函数 l 在 $\mathbb{R}^{H\times H}\times\mathbb{R}^{H\times H}$ 上连续. 那么应用连续映射定理得 $\mathbb{P}$-a.s.，

$$l(\hat{\vec{\mu}}_t(n), \hat{J}(n)) \to l(\vec{\mu}_t, J),\ n \to \infty,$$

即有 $\mathbb{P}$-a.s.，

$$\hat{\vec{\mu}}_T(n) \to \vec{\mu}_0 e^{TQ} J,\ n \to \infty.$$

进一步，由有界收敛定理得

$$E[\hat{\vec{\mu}}_T(n)] \to \vec{\mu}_0 e^{TQ} J,\ n \to \infty,$$

即证得 $\hat{\vec{\mu}}_T(n)$ 是 $\vec{\mu}_T$ 的渐进无偏估计.

2.6 多阶段马氏链模型概率分布预测

在前面的讨论中，本文考虑的是在时间 $[0,T]$ 内的马氏链数学模型，它是一阶段多状态转移的数学模型，接下来考虑多阶段的多状态转移问题. 现在对于 $n \geqslant 1$，定义随机过程 $X'=\{X'_t\}_{t\in[0,T_n]}$，其包含 n 个阶段，$X^1=\{X'_t\}_{t\in[0,T_1]}$，$X^2=\{X'_t\}_{t\in[T_1,T_2]}$，…，$X^n=\{X'_t\}_{t\in[T_{n-1},T_n]}$，假设每个阶段都服从本文前面讨论的一阶段的模型. 设 $\vec{\upsilon}_t$ 代表随机变量 X'_t 的分布，其中 $t \in [0,T_n]$. 进一步，对第 $i(1 \leqslant i \leqslant n)$ 个阶段进行参数估计可以得到强度矩阵和跳跃矩阵的估计 $\hat{Q}_i$ 和 $\hat{J}_i(1 \leqslant i \leqslant$

n). 现在考虑第 $i(1<i\leqslant n)$ 个阶段，即 $t\in[T_{i-1},T_i]$，显然第 i 个阶段的初始分布的估计为 $\hat{\boldsymbol{\upsilon}}_{T_{i-1}}$，则应用定理 2.4 可知当 $i=2,3,\cdots,n$ 时，t 时刻分布的估计为

$$\left.\begin{aligned}\hat{\boldsymbol{\upsilon}}_t&=\hat{\boldsymbol{\upsilon}}_{T_{i-1}}\mathrm{e}^{(t-T_{i-1})\hat{Q}_i},\ T_{i-1}<t<T_i;\\ \hat{\boldsymbol{\upsilon}}_{T_i}&=\hat{\boldsymbol{\upsilon}}_{T_{i-1}}\mathrm{e}^{(T_i-T_{i-1})\hat{Q}_i}\hat{J}_i,\ t=T_i\end{aligned}\right\}\tag{4}$$

3　实证分析

为了检测本文马氏链模型的预测效果，我们将国内某寿险公司六种保险的保单数据应用于本文的马氏链模型中，数据来源于该寿险公司的个人保单信息原始数据. 险种 1～6 分别为分红终生寿险、定价利率为 7.5%的终生寿险、定价利率为 2.5%的终生寿险、定价利率为 2.5%的两全保险、定价利率为 5.0%的两全保险、定价利率为 5.0%的终生寿险. 此次实证中分段考虑的时间间隔为 1 年，由于失效行为只能在保单年度末被保险公司发现，故将跳跃矩阵 J 产生的时刻选为保单年度末. 实际上，在保单到达满期之前可能发生如图 1 所示的状态转移，箭头表示转移方向.

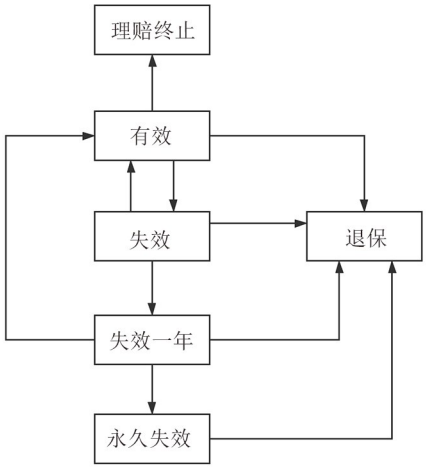

图 1　保单状态转移情况

Fig. 1　State transition of insurance policy

用数字 1～6 分别表示保单状态有效、失效、失效 1 年、永久失效、退保和理赔终止. 则状态集合 $\mathscr{H}=\{1,2,3,4,5,6\}$. 由图 1 可知强度矩阵 Q 和跳跃矩阵 J 可分别表示为

$$Q=\begin{bmatrix}q_{11}&q_{12}&0&0&q_{15}&q_{16}\\q_{21}&q_{22}&q_{23}&0&q_{25}&0\\q_{31}&0&q_{33}&q_{34}&q_{35}&0\\0&0&0&q_{44}&q_{45}&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix},$$

$$J=\begin{bmatrix}j_{11}&j_{12}&0&0&j_{15}&j_{16}\\j_{21}&j_{22}&j_{23}&0&j_{25}&0\\j_{31}&0&j_{33}&j_{34}&j_{35}&0\\0&0&0&j_{44}&j_{45}&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{bmatrix}.$$

由于失效行为只能在保单年度末才能被保险公司发现，而强度矩阵 Q 所在的时间范围为 $[0,T)$，因此 q_{12},q_{23},q_{34} 为 0. 再利用强度矩阵对角元素 $q_{ii}=-\sum\limits_{j\in\mathscr{H},j\neq i}q_{ij}$，那么对于强度矩阵 Q 需要估计的元素为 $q_{15},q_{16},q_{21},q_{25},q_{31},q_{35}$ 和 q_{45}.

同样的，由于保单年度末为跳跃矩阵 J 产生的时刻，因此 $j_{15},j_{16},j_{21},j_{25},j_{31},j_{35}$ 和 j_{45} 均为 0，由失效状态转为失效 1 年的概率为 1，即 $j_{23}=1$. 同样的由失效 1 年转为永久失效状态的概率也是 1，即 $j_{34}=1$. 再由 $j_{ii}=1-\sum\limits_{k\in\mathscr{H},k\neq i}j_{ik}$ 可知对于跳跃矩阵 J 需要估计的元素为 j_{12}.

现在利用实际数据来对 $q_{15},q_{16},q_{21},q_{25},q_{31},q_{35},q_{45}$ 和 j_{12} 进行估计. 在固定样本的前提下，利用数据可以得到在 $n(n\in\mathbb{N})$ 次随机抽样中由状态 $i(i\in\mathscr{H})$ 转移至状态 $j(j\in\mathscr{H},j\neq i)$ 的总次数 $\sum\limits_{k=1}^{n}m^k(i,j)$ 和 n 次抽样中到达状态 i 时停留的总时长 $\sum\limits_{k=1}^{n}\left(\sum\limits_{\alpha=1}^{m^k(i)}t_\alpha^k(i)\right)$，利用式(1)即得强度矩阵 Q 的非对角元素 $q_{ij}(i\neq j)$ 的估计值，再利用 $q_{ii}=-\sum\limits_{j\in\mathscr{H},j\neq i}q_{ij}$ 可以得到 q_{ii} 的估计值. 同样可以得到在 T 时刻的状态为 $i(i\in\mathscr{H})$ 的样本数 $n(i)$，和在 T 时刻由状态 i 转移至状态 $k(k\in\mathscr{H})$ 的样本数 $n(i,k)$，利用式(2)可得跳跃矩阵 J 的元素 j_{ik} 的估计值，即得 j_{12} 的估计值，再由 $j_{11}=1-j_{12}$ 可以得到 j_{11} 的估计值. 利用式(4)可以得到各阶段的概率分布的预测，即可以得到保单在生效之后每年转变成退保或者失效的概率. 将得到的预测值与实际值进行对比，得到如图 2 所示的结果. 由图 2 可看出本文模型有较好的预测效果.

为了更加精准地分析此模型的预测效果，现在对其进行误差分析，计算误差的公式为

$$\text{误差}=|\text{预测值}-\text{真实值}|/\text{真实值}.$$

由于在本次实证中需要估计的是保单生效后每年的强度矩阵和跳跃矩阵，因此实证中的预测误差也主要是在估计强度矩阵和跳跃矩阵时产生的. 在马氏链模型的假设条件中，我们假设在时间 $[0,T)$ 内保单状态服从连续时间时齐马氏链，因此它的强度矩阵 Q 为常数矩阵，而在实证中将 T 取为保单年度末，即在一整年的时间里，马氏链模型的强度矩阵 Q 都为一个常数，因此会对预测值产生一定的影响. 跳跃矩阵 J 是在保单年度末产生的，其代表着在保单年度末保单状态转移情况的概率转移矩阵，在样本足够大的情况下得到的估计值与真实值相近，因此跳跃矩阵 J 的估计也会对预测值产生影响.

我们利用本文的马氏链模型和数据得到 107 个误差值，结果如表 1 所示. 由表 1 可以发现误差较大(误差值大于 7%)的预测值共有 6 个，占全部预测值的 5.6%，可见此模型在实际中能起到较好的预测效果.