第47卷第5期
2007年9月
大连理工大学学报
Journal of Dalian University of Technology
Vol. 47, No. 5
Sept. 2007

文章编号: 1000-8608(2007)05-0777-04

预测死亡率分布的一个最小叉熵模型

姜昱汐*[1,2], 李兴斯[3], 李 华[4]

(1. 大连理工大学 管理学院, 辽宁 大连 116024;
2. 大连交通大学 管理学院, 辽宁 大连 116028;
3. 大连理工大学 工业装备结构分析国家重点实验室, 辽宁 大连 116024;
4. 鞍山科技大学 经济管理学院, 辽宁 鞍山 114044)

摘要: 被保人的死亡率分布是确定寿险费率的一个重要依据, 而根据其生活的环境、时间预测被保人的死亡率是保险精算研究中的一个热点问题. 基于最小叉熵原理, 建立了预测被保人死亡率分布的一个模型——最小叉熵模型, 该模型以叉熵函数作为目标函数, 以被保人的预期寿命作为约束条件, 通过最小化叉熵预测被保人的死亡率. 以从事特殊职业的被保人为研究对象, 通过最小叉熵模型计算了该类人的死亡率. 该方法计算简便, 具有较好的客观性和实用性, 为死亡率预测研究提供了一种有效的新方法.

关键词: 死亡率; 寿险保费; 最小叉熵原理
中图分类号: O221.2 文献标识码: A

0 引 言

人寿保险以人的寿命为保险标的, 以被保人的生存与死亡决定保险金的给付与否和给付时间. 因此, 被保人的生命规律对寿险经营有着直接的影响. 而且, 在人寿保险中, 从保险费率的厘定、责任准备金的提取、保单现金价值的计算到保单红利的分配, 都必须考虑一个重要的因素——死亡率, 各年龄的死亡率构成一个生命表. 在保险工作中, 常常需要对一些特殊人群进行担保, 例如针对在某些特殊环境下工作的人的保险, 针对某些曾经患过某种疾病的人的保险等, 这些人的死亡率会因为他们自身的特殊条件而不符合标准的生命表. 如何根据被保人的具体情况给出适合该类人的死亡率分布, 制定出恰当的保费是保险精算中常常需要面临的一个问题. 而且近年来, 随着人们生活质量和医疗水平的提高, 人类整体的平均寿命有所增加, 使得被保人在未来的真实死亡率与根据经验数据统计得到的死亡率不同, 这就需要根据被保人的寿命特征对死亡率进行预测, 重新修订生命表, 进而收取与未来死亡率相对应的保费, 这对寿险公司的运营安全至关重要.

考查现有预测死亡率的模型[1-3], 它们大都建立在两个基本条件上: 一个是现有的经验生命表; 一个是专家对被保人寿命的调查研究. 但是这些模型或是计算不方便, 或是得到的预测结果不理想.

本文在前人研究的基础上, 同样基于上述两个条件, 以最小叉熵原理作为统计推断工具, 建立最小叉熵模型, 并将该模型应用于预测从事危险职业人群的死亡率分布.

1 最小叉熵原理及其数学表述

最小叉熵原理提供了一个求概率分布的准则[4], 即若想使待求的概率分布在满足一定的约束条件下, 尽可能地靠近一个已知的概率分布, 则必须令叉熵函数极小化. 这个原理在数学上可表达为如下的数学规划问题:

$$
\begin{aligned}
&\min D(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{n} p_i \ln(p_i/q_i) \\
&\text{s.t.} \sum_{i=1}^{n} p_i g_j(x_i) = E[g_j],\ j = 1,2,\cdots,m \\
&\qquad \sum_{i=1}^{n} p_i = 1 \\
&\qquad (p_i > 0,\ i = 1,2,\cdots,n) \qquad (1)
\end{aligned}
$$

其中 $\boldsymbol{p} = (p_1\ \ p_2\ \ \cdots\ \ p_n)^{\mathrm{T}}$ 为待估的概率分布,

收稿日期: 2006-01-08; 修回日期: 2007-07-23.
基金项目: 国家自然科学基金资助项目(重大项目 10590354); 国家自然科学基金资助项目(10572031).
作者简介: 姜昱汐* (1975-), 女, 博士, 讲师; 李兴斯 (1942-), 男, 教授, 博士生导师.

是优化问题的变量; $\boldsymbol{q} = (q_1\quad q_2\quad \cdots\quad q_n)^{\mathrm{T}}$ 为已知的概率分布 $(\sum_{i=1}^{n} q_i = 1$,且 $q_i > 0, i = 1, 2,\cdots, n)$,在优化过程中作为已知参数, $g_j\,(j = 1, 2,\cdots, m)$ 表示各阶统计矩函数, $E[\ ]$是由实验观测得到的各阶统计矩的期望值.

在上述优化问题中欲最小化的目标函数 $D(\boldsymbol{p},\boldsymbol{q})$,即所谓的叉熵函数,代表由待求概率分布 $\boldsymbol{p}$到已知概率分布 $\boldsymbol{q}$的单向"距离". 文献中一般将其称为 Kullback-Leibler散度或方向散度[4]. 需注意的是,这个距离与一般的距离定义不同,其一它是带方向的,其二它不满足一般距离定义中的三角不等式关系. 容易验证,该函数具有如下几个重要性质:

(1) $D(\boldsymbol{p},\boldsymbol{q})$是一个非负的、非对称的连续函数;

(2) $D(\boldsymbol{p},\boldsymbol{q})$是 $p_i\,(i = 1, 2,\cdots, n)$的凸函数;

(3) 当且仅当 $\boldsymbol{p} = \boldsymbol{q}$时, $D(\boldsymbol{p},\boldsymbol{q}) = 0$.

现在,考查一下上述优化问题. 若问题(1)中没有统计矩约束,那么求得的概率分布必然为已知的概率分布,即 $\boldsymbol{p} = \boldsymbol{q}$. 矩约束的参与,将使求得的概率分布偏离已知分布. 然而,这个优化问题的目标,即最小化函数 $D(\boldsymbol{p},\boldsymbol{q})$,起到使这种偏离尽可能小的作用. 因此,根据最小叉熵原理所求得的概率分布,是在服从已知信息(矩约束)下的最接近已知分布的一组概率分布.

因为距离函数 $D(\boldsymbol{p},\boldsymbol{q})$为凸函数,并且问题(1)里的约束函数为线性函数,因此该优化问题为凸规划. 根据数学规划的对偶理论,应存在相应的对偶规划. 而且由于上述优化问题是一个变量可分离的凸规划问题,很容易通过拉格朗日函数的驻值条件,求得其封闭形式的全局最优解.

构造问题(1)的拉格朗日函数

$$L(\boldsymbol{p},\boldsymbol{q},\lambda) = \sum_{i=1}^{n} p_i \ln(p_i/q_i) - (\lambda_0 + 1)\times \left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{j=1}^{m}\lambda_j\left(\sum_{i=1}^{n} p_i g_j(x_i) - E[g_j]\right) \tag{2}$$

由驻值条件和归一化条件求得问题(1)的全局最优解为

$$p_i = q_i \exp\left[\sum_{j=1}^{m}\lambda_j g_j(x_i)\right] \Big\backslash Z_q,\ i = 1, 2,\cdots, n \tag{3}$$

其中

$$Z_q = \sum_{i=1}^{n} q_i \exp\left[\sum_{j=1}^{m}\lambda_j g_j(x_i)\right] \tag{4}$$

将式(3)给出的解代入式(2)的 $L(\boldsymbol{p},\boldsymbol{q},\lambda)$中,消去 $\boldsymbol{p}$,即可得到最小叉熵问题的对偶规划为

$$\max_{\lambda\in\mathbf{R}^m} D_q(\lambda) = -\ln Z_q + \sum_{j=1}^{m}\lambda_j E[g_j] \tag{5}$$

由于这个对偶规划是一个无约束优化问题,并且对偶变量数仅为原问题的矩约束个数 m,通常要比原变量数 n小得多,解相应的对偶规划问题要比直接解原问题容易得多,而且还可以直接利用现成的解无约束优化问题的软件求解. 因此对于叉熵优化问题,一般应解相应的对偶规划.

2 最小叉熵模型在预测死亡率上的应用

在人寿保险中,有一类弱体保险,该保险是针对身体有缺陷的人或从事某种危险职业的人,该类人的死亡率高于经验生命表中给出的死亡率,因此其平均余寿必然低于经验生命表中同龄人的平均余寿[5].

本文在求这类特殊人群的死亡率分布时,以经验生命表中的死亡率分布作为基准,以满足特殊人群的统计矩为约束条件,根据最小叉熵原理来获得特殊人群的死亡率分布.

下面以从事某类危险职业(如从事石棉作业)的人群为例,介绍通过最小叉熵模型预测死亡率的方法. 为简单计,仅以取整平均余寿 c作为约束条件,建立求解该类人群死亡率分布的数学模型如下:

$$\begin{aligned}
&\min D(\tilde{\boldsymbol{p}},\tilde{\boldsymbol{q}}) = \sum_{i=0}^{n}\tilde{p}_i \ln(\tilde{p}_i/\tilde{q}_i)\\
&\text{s. t.}\ \sum_{i=0}^{n} i\tilde{p}_i = c\\
&\qquad \sum_{i=0}^{n}\tilde{p}_i = 1\\
&\qquad\qquad (\tilde{p}_i > 0,\ i = 0, 1, 2,\cdots, n)
\end{aligned} \tag{6}$$

其中波浪符号代表归一化概率 $\tilde{p}_i = d^p_{x+i}/l_x$ 和 $\tilde{q}_i = d^e_{x+i}/l_x$,它们分别为 x岁的人在 $x+i$岁到 $x+i+1$岁之间的死亡人数预测值 d^p_{x+i}(经验值 d^e_{x+i})除以在 x岁(即初始时)的生存人数 l_x计算出来的概率.

求解式(6),通过拉格朗日函数的驻值条件,可以求得

$$\tilde{p}_i = \tilde{q}_i \exp(i\lambda)\Big\backslash\sum_{i=0}^{n}\tilde{q}_i \exp(i\lambda),\ i = 0, 1, 2,\cdots, n \tag{7}$$

则其对偶规划为

$$\max_{\lambda\in\mathbf{R}} D_q(\lambda) = -\ln\sum_{i=0}^{n}\tilde{q}_i \exp(i\lambda) + c\lambda \tag{8}$$

求解式(8),将解出的拉格朗日乘子 _ 代回式(7),即可得到所求. 鉴于在本文的示例中仅考虑取整平均余寿作为约束条件,只需解一个单变量 _ 的无约束优化问题(8),即可求出概率分布 $\tilde{\boldsymbol{p}}$,然后通过简单变换得到与经验生命表中对应的死亡率,计算非常简便容易.

根据上述的最小叉熵模型,以年龄为 60~105岁 (男性)为被保人,即起始年龄 $x=60$, $n=45$,以 1990~1993年中国人寿保险业的经验生命表为基准. 假设从事某种危险职业人群的取整平均余寿 $c=12$(经验生命表中 60岁男性的平均余寿为 18.79,取整平均余寿为 18.29[6]),根据该模型来预测 $\tilde{p}_i$.

计算结果见表 1. 表 1中 $p_i = d^p_{x+i}/l^p_{x+i}$,为预测得到的死亡率 (即与经验生命表中对应的死亡率),其中 $d^p_{x+i} = \tilde{p}_i l_x$,为 x岁的人在 $x+i$岁到 $x+i+1$岁之间的死亡人数的预测值; $l^p_{x+i} = l^p_{x+i-1} - d^p_{x+i-1}$,为 x岁的人在 $x+i$岁时的生存人数的预测值.

表 1　死亡率的经验值与预测值

Tab. 1　The mortality of standard and forecasting

年龄	经验值				预测值			
	q_i	$\tilde{q}_i$	l^q_{x+i}	d^q_{x+i}	p_i	$\tilde{p}_i$	l^p_{x+i}	d^p_{x+i}
60	0.013 553	0.013 553	853 391	11 566	0.044 954	0.044 954	853 391	38 363
61	0.014 892	0.014 690	841 825	12 536	0.047 110	0.044 992	815 027	38 396
62	0.016 361	0.015 899	829 289	13 568	0.049 410	0.044 966	776 631	38 374
63	0.017 972	0.017 178	815 721	14 660	0.051 860	0.044 864	738 258	38 286
64	0.019 740	0.018 530	801 061	15 813	0.054 480	0.044 686	699 971	38 134
65	0.021 677	0.019 946	785 248	17 022	0.057 274	0.044 418	661 837	37 906
66	0.023 800	0.021 425	768 226	18 284	0.060 259	0.044 057	623 931	37 597
67	0.026 125	0.022 958	749 942	19 592	0.063 447	0.043 592	586 334	37 201
68	0.028 671	0.024 537	730 350	20 940	0.066 861	0.043 023	549 132	36 715
69	0.031 457	0.026 150	709 410	22 316	0.070 511	0.042 338	512 417	36 131
70	0.034 504	0.027 780	687 094	23 707	0.074 416	0.041 532	476 286	35 443
71	0.037 835	0.029 411	663 387	25 099	0.078 600	0.040 603	440 843	34 650
72	0.041 474	0.031 020	638 288	26 472	0.083 080	0.039 544	406 193	33 746
73	0.045 446	0.032 582	611 816	27 805	0.087 880	0.038 354	372 446	32 731
74	0.049 779	0.034 065	584 011	29 071	0.093 018	0.037 029	339 715	31 600
75	0.054 501	0.035 441	554 940	30 245	0.098 527	0.035 573	308 116	30 358
76	0.059 644	0.036 671	524 695	31 295	0.104 428	0.033 989	277 758	29 006
77	0.065 238	0.037 718	493 400	32 188	0.110 746	0.032 281	248 752	27 548
78	0.071 317	0.038 543	461 212	32 892	0.117 514	0.030 460	221 204	25 995
79	0.077 916	0.039 106	428 320	33 373	0.124 762	0.028 539	195 209	24 355
80	0.085 069	0.039 370	394 947	33 598	0.132 515	0.026 530	170 855	22 641
81	0.092 813	0.039 300	361 349	33 538	0.140 805	0.024 455	148 214	20 869
82	0.101 184	0.038 867	327 811	33 169	0.149 663	0.022 333	127 345	19 059
83	0.110 218	0.038 054	294 642	32 475	0.159 123	0.020 191	108 286	17 231
84	0.119 951	0.036 849	262 167	31 447	0.169 208	0.018 054	91 055	15 407
85	0.130 418	0.035 259	230 720	30 090	0.179 955	0.015 952	75 648	13 613
86	0.141 651	0.033 301	200 630	28 419	0.191 384	0.013 912	62 035	11 872
87	0.153 681	0.031 012	172 211	26 465	0.203 525	0.011 963	50 162	10 209
88	0.166 534	0.028 441	145 746	24 271	0.216 399	0.010 131	39 953	8 646
89	0.180 233	0.025 654	121 475	21 893	0.230 022	0.008 438	31 307	7 201
90	0.194 795	0.022 730	99 582	19 398	0.244 419	0.006 904	24 106	5 892

(续表)

年龄	经验值				预测值			
	q_i	$\tilde{q}_i$	l^q_{x+i}	d^q_{x+i}	p_i	$\tilde{p}_i$	l^p_{x+i}	d^p_{x+i}
91	0.210 233	0.019 753	80 184	16 857	0.259 579	0.005 540	18 214	4 728
92	0.226 550	0.016 811	63 327	14 346	0.275 507	0.004 354	13 486	3 715
93	0.243 742	0.013 989	48 981	11 938	0.292 208	0.003 345	9 770	2 855
94	0.261 797	0.011 363	37 043	9 697	0.309 659	0.002 509	6 915	2 141
95	0.280 694	0.008 994	27 346	7 675	0.327 833	0.001 834	4 774	1 565
96	0.300 399	0.006 923	19 671	5 908	0.346 681	0.001 304	3 209	1 112
97	0.320 871	0.005 173	13 763	4 415	0.366 173	0.000 900	2 096	768
98	0.342 055	0.003 745	9 348	3 196	0.386 175	0.000 601	1 329	513
99	0.363 889	0.002 621	6 152	2 237	0.406 623	0.000 389	816	332
100	0.386 299	0.001 770	3 915	1 511	0.427 417	0.000 242	484	207
101	0.409 200	0.001 151	2 404	982	0.447 973	0.000 145	277	124
102	0.432 503	0.000 718	1 422	613	0.467 771	0.000 084	153	71
103	0.456 108	0.000 430	809	367	0.485 883	0.000 046	82	40
104	0.479 911	0.000 246	442	210	0.499 362	0.000 025	42	21
105	1.000 000	0.000 267	232	228	1.000 000	0.000 024	21	21

3 结 语

本文将最小叉熵模型运用于预测死亡率分布,通过求解最小叉熵模型的对偶规划,以特殊被保人的取整平均余寿作为约束条件,求出该类人的死亡率分布,为确定适合弱体保险的死亡率分布提供了一种简便有效的方法. 该模型计算容易,具有一定的实用价值.

考虑到我国的经验生命表都是 10年前的,而随着生活水平的提高,我国国民的寿命也有所增长,用现有的经验生命表来制定保费并不符合实际,而如果通过人口普查重新编订生命表的工作量实在太大,使用本文提出的最小叉熵模型,根据专家对国民平均寿命增长的预估值预测被保人未来的死亡率,对经验生命表进行与时俱进的调整,可为寿险费率厘定提供更为合理的依据.

参考文献:

[1] LEE R, CARTER L. Modeling and forecasting U. S. mortality [J]. **J Amer Stat Assoc**, 1992, **87**(419): 659-671

[2] ALHO J, SPENCER B. Uncertain population forecasting [J]. **J Amer Stat Assoc**, 1985, **80**(390): 306-314

[3] LEE R. The Lee-Carter method for forecasting mortality, with various extensions and applications [J]. **North Amer Actuarial J**, 2000, **4**(1): 80-93

[4] KULLBACK S. **Information Theory and Statistics** [M]. New York: John Wiley, 1959

[5] 王铁豪,徐福君. 现代保险学 [M]. 青岛: 青岛出版社,1995

[6] 曾庆五,陈迪红,黄大庆. 保险精算技术 [M]. 大连: 东北财经大学出版社,2002

A minimum cross entropy model for forecasting mortality

JIANG Yu-xi* [1,2], LI Xing-si [3], LI Hua [4]

(1. School of Manage., Dalian Univ. of Technol., Dalian 116024, China;
2. School of Manage., Dalian Jiaotong Univ., Dalian 116028, China;
3. State Key Lab. of Struct. Anal. for Ind. Equip., Dalian Univ. of Technol., Dalian 116024, China;
4. School of Econ. and Manage., Anshan Univ. of Sci. and Technol., Anshan 114044, China)

Abstract Mortality distribution is an important criterion to determine the ratemaking of life insurance, and it has been attracting the significant interest of researchers to forecast the mortality of people living in different environments or periods in recent years. Based on the minimum cross entropy theory, a new model, which can forecast mortality distribution by minimizing the cross entropy function subject to the future life expectancy of insurants, is proposed. In addition, the mortality of the people working in the particular environment is calculated. From the process it can be seen that the model can be easily solved by simple calculation, and it is efficient.

Key words mortality; premium of life insurance; minimum cross entropy theory